POLITECNICO DI TORINO

POLITECNICO DI TORINO

Master of Science's Degree in Data Science and Engineering

Master of Science's Degree Thesis

Predictive Model For Humanitarian Aid: Research on a conflict early warning system for the Sahel region.

Supervisor
Prof. Giuseppe Rizzo

Candidate
Luca Varriale

Academic Year 2023-2024

Summary

This study explores the complex situation in the Sahel region, an area characterized by continuous social conflicts and significant political instability. It presents an early warning system developed to predict violence across 12 Sahelian countries, focusing on the urgent need for accurate forecasting to guide policy decisions and predict catastrophic events. Utilizing historical data, the system predicts fatalities at both national and regional level, in order to enable policymakers and stakeholders to take appropriate action in a timely manner, given the expected conflict.

Central to the study is the integration of data from multiple sources. In particular, *GDELT* provides structured information extracted from online news articles relevant to regional conflict dynamics, while *ACLED* offers a comprehensive dataset on political violence, protests, and associated events. One of the main challenges in conflict early warning systems is their limited ability to generalize across different conflict intensities, since the country's historical data mainly represents peaceful periods, making it difficult for models to accurately predict potential escalations into conflict. To address this problem, synthetic data is generated through state-of-the-art data augmentation methods to enable the model to be more adaptive to changes in scenarios, including wartime and peacetime contexts.

Model calibration is one of the critical elements in this predictive framework, ensuring that confidence levels associated with the forecasts actually represent the true likelihood of future events. Through the application of calibration techniques, such as conformal prediction, the model accounts for the unique temporal dynamics and potential covariate shifts inherent in the data, improving the reliability of predictions.

Another consideration is the ethical responsibility to deploy predictive models related to conflict-prone countries. This research emphasizes data minimization whenever possible and encourages reflection on the broader societal implications of predictive analytics. Addressing those technical and ethical dimensions, this research aims to contribute to an effective and humane approach in managing and mitigating the conflicts in the Sahel region.

Contents

List of Figures

List of Tables

Acronyms

ACLED	Armed Conflict Location and Event Data Project
AU	African Union
CAMEO	Conflict and Mediation Event Observations
CEWARN	Conflict Early Warning and Response Mechanism
CEWS	Continental Early Warning System
ECOWARN	ECOWAS Early Warning and Response Network
ECOWAS	Economic Community of West African States
EWS	Early Warning Systems
EFB	Exclusive Feature Bundling
GCRI	Global Conflict Risk Index
GDELT	Global Database of Events, Language, and Tone
GDPR	General Data Protection Regulations
GOSS	Gradient-based One-Side Sampling
GTI	Global Terrorism Index
GWS	Green Water Scarcity Index
HDR	Human Development Reports
HDI	Human Development Index
IGAD	Intergovernmental Authority on Development
LightGBM	Light Gradient Boosting Machine
KDE	Kernel Density Estimation
NNI	Natural Neighbor Interpolation
PRIO-GRID	Peace Research Institute Oslo Grid
UNDP	United Nations Development Programme
XG-BOOST	eXtreme Gradient Boosting
ViEWS	Violence Impacts Early-Warning System
WPS	Water, Peace, and Security

Chapter 1

Introduction

Conflicts continue to pose significant threats to human life and societal stability across various regions of the world. The Sahel region, in particular, stands as one of the most critical areas, characterized by persistent social conflicts and substantial political instability. In this complex scenario, the development of effective early warning systems becomes of prime importance for preventing catastrophic events and guiding policy decisions.

This thesis presents a comprehensive early warning system designed to predict violence outbursts across 12 Sahelian countries. The system leverages different models to forecast conflicts, integrate causal inference methods with data augmentation techniques and enable the model to adapt to diverse scenarios ranging from peacetime to conflict situations.

The central key to our methodology is the fusion of multiple data sources. We combine structured information, particularly from GDELT, extracted from online news articles relevant to regional conflict dynamics, and ACLED's comprehensive dataset on political violence, protests, and associated events.

Model calibration represents another important aspect of our framework. Through the application of conformal prediction techniques, we ensure that the confidence levels associated with our forecasts accurately represent the true likelihood of future events.

1.1 Main Contributions

The structure of the thesis is organized as follows:

- In Chapter 2, we present the Sahel region, in order to gain a deeper knowledge of the difficulties it faces.
- In Chapter 3, we will analyze the state of the art conflict early warning systems and discuss some ethical considerations related to them.
- In Chapter 4, we provide a detailed overview of the datasets analyzed and used in this work.
- In Chapter 5, we will report the analytical steps through which our collected data can be converted into the final features considered by the model.

- In Chapter 6, we will explore our experimental framework that led to our final results, presenting the employed models, causal inference methods, probability distribution estimation techniques, data augmentation approaches, model calibration strategies and our obtained results.
- In Chapter 7, we provide the conclusions and future work.

Chapter 2

The Sahel Region

In this chapter, we aim to frame the Sahel region and its different aspects to gain a deeper knowledge of the difficulties it faces. First, we will explore the geographical and environmental context of the Sahel, which is essential for comprehending the dynamics and issues that characterize this area. Subsequently, we will consider the urgent need for effective conflict management in the Sahel, examining the various obstacles currently at play, including terrorism and conflicts, water scarcity, overpopulation, climate change and migration, as well as poverty, economic instability, and political fragility. This examination establishes a foundation for contextualizing the data that we will analyze later, as it provides a fundamental insight into the underlying factors influencing the region's instability.

2.1 Geography and Ecology of the Sahel

The Sahel, represented in Fig. 2.1, is a vast transitional zone that extends across the southern latitudes of North Africa, forming a band between the Sahara Desert to the north and savannas to the south [43]. This region stretches approximately 5,900 kilometers from the Atlantic Ocean to the Red Sea, with a width of several hundred kilometers between 12°N and 20°N latitudes [24]. Topographically, the Sahel is quite flat; the elevation varies between 200 and 400 meters, with some isolated mountain ranges like the Aïr Mountains and Marrah Mountains rising out of the surrounding plains. The Sahel is characterized by arid and semi-arid landscapes that are dominated by open grasslands with minimal arboreal vegetation.

This area covers a space of 3 million square kilometers, crossing over twelve countries which include Senegal, Mauritania, Mali, Burkina Faso, Niger, Nigeria, Cameroon, Chad, the Central African Republic, South Sudan, Sudan and Eritrea [43]. The Sahel region is within the tropical latitude but characterized by hot semi-arid and not equatorial climate. Acting as a geographical and ecological boundary, the region plays a crucial role in shaping the environmental patterns of North Africa.

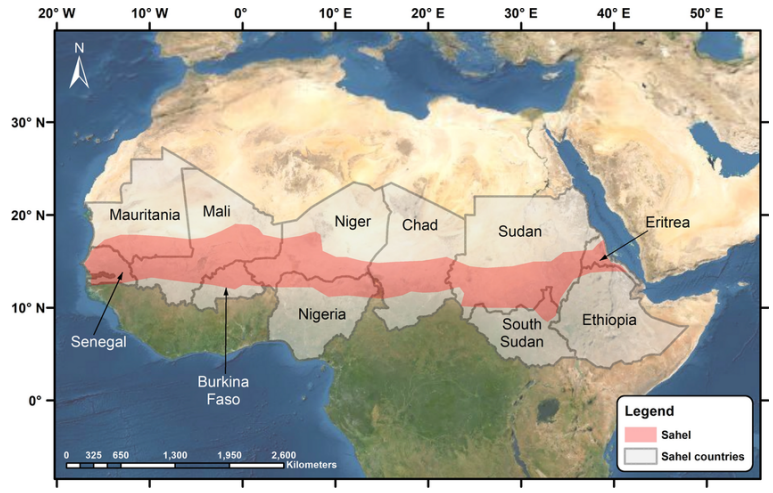

Figure 2.1. The red stripe represents the Sahel region [5].

2.2 The Need for Conflict Management in the Sahel

The Sahel region is experiencing an escalating crisis marked by violence, political competition, and widespread insecurity. The terrorist threat from Boko Haram, Al-Qaeda, and Islamic State networks has intensified throughout the region, and significant numbers of atrocities have been committed. Such acts have come at a great cost as developing countries' economies are at risk while people's livelihoods are heavily impacted and millions are internally-displaced. Furthermore, terrorist organizations are now infiltrating criminal networks as well as existing in areas where human trafficking has become the crime of choice, which can have further implications with regard to the undermining of governmental authority and the destabilization of social structures. This security deterioration crosses national boundaries, with countries such as Mali, Burkina Faso, Niger, and Nigeria experiencing significant internal conflicts that weaken their political foundations.

The impact of this instability is severe. Regional governments find themselves forced to increase military expenditure, reducing social service funding, as exemplified by Mali's fourfold increase in defense spending over five years. This shift toward military priorities strains already limited resources, deepening these nations' economic difficulties and their reliance on international support.

These circumstances constrain the Sahelian states' capacity to prevent episodes of violence. Traditional conflict prevention and mitigation approaches often respond too slowly and suffer from inadequate data. The Sahel region requires effective conflict management tools. Without targeted intervention, the pattern of violence, economic disruption, and social fragmentation will continue to expand.

2.3 Sahel Challenges

This section presents a general overview of the interconnected crises that provide an understanding of the complex nature of the Sahel's challenges and their implications for the stability and security of the region. Some of the complex challenges in the Sahel include terrorism, organized crime, extreme lack of access to water, and migration driven by climatic changes. These are aggravated by rapid population growth, lack of economic opportunities, and political instability, resulting in a self-intensifying pattern. We will consider how terrorism and violent extremism have spread throughout the Sahel to such an extent that this are has become an epicenter of terrorist activity and organized crime. Another major driver of conflict in this region is water scarcity, which could be accentuated by drought and pressure from the environment, heightening tensions over basic resources. The discussion then expands to the overpopulation pressures that stretch the already scare Sahel resources, which further complicates socio-economic stability. These challenges are also worsened by the strong impacts of climate change, where rising temperatures and erratic rainfall patterns have pushed many people to migrate, putting significant strain on the fragile ecosystem. Adding to these difficulties are widespread poverty, economic instability, and weak governance, which together prevent Sahelian countries from effectively addressing these pressing challenges.

2.3.1 Terrorism and Conflicts

In recent years, the Sahel region has been identified as the epicenter of global terrorism with almost a half of all related deaths emanating from this region. According to the global terrorism index 2024 [3], this increase is mainly attributed to relocation by terrorists from Middle East states into parts of Sub-Sahara Africa especially after ISIS and other extremist organizations got defeated in those areas around the year 2019. Figure 2.2 clearly shows that terrorist attacks have been increasing gradually since 2007 up to 2023, with the decline of terrorists' influence in the Middle East. In 2023, for the first time, Burkina Faso emerged top on the Global Terrorism Index (GTI) among Sub Saharan nations showing an increase of seventy percent on terrorism related deaths within one year [3].

Additionally, terrorism in the Sahel is closely linked with organized crime. Figure 2.3 illustrates the strong correlation between the impact of terrorism and the prevalence of organized criminal activities. This relationship is particularly strong in regions like the Sahel, where weak governance allows both terrorist groups and criminal networks to proliferate. As shown, the Sahel exhibits a high correlation coefficient ($r = 0.81$) between the Global Terrorism Index and the Organized Crime Index, indicating the extent to which these factors are correlated [3]. Terrorist groups often benefit from illegal activities, such as cattle rustling, gold mining, and drug trafficking, while using violence to control or expand their influence in regions where state presence is weak.

As indicated in 2.4 countries with low Positive Peace are susceptible to additional ecological dangers. [21] defines Positive Peace as attitudes, institutions and structures that create and maintain peaceful societies. The level of socio-economic development within a given society is measured; its resiliency, its future potential economic growth,

and its ability to address conflicts or tensions within its population without resorting to violence.

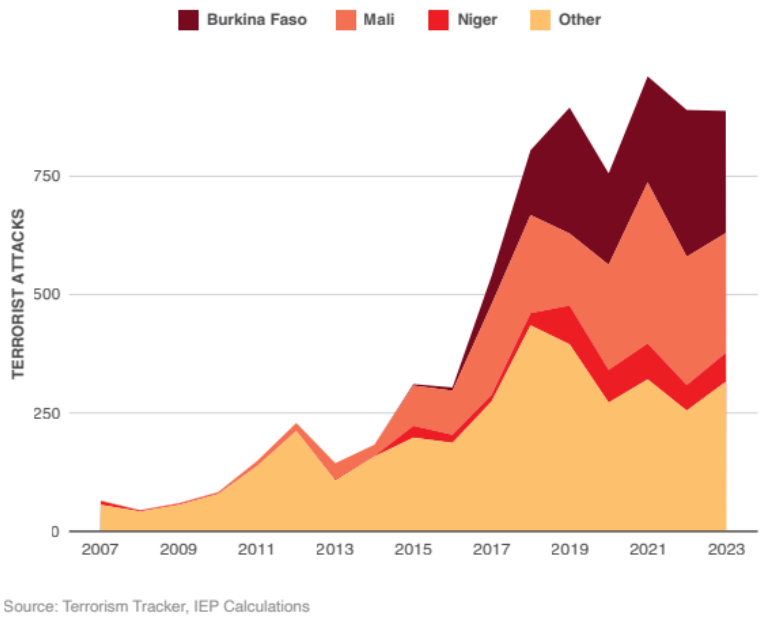

Figure 2.2. Terrorist Attacks in the Sahel, 2007-2023. The graph illustrates how the decline of ISIS in the Middle East coincided with the surge in terrorist activity in the Sahel [3].

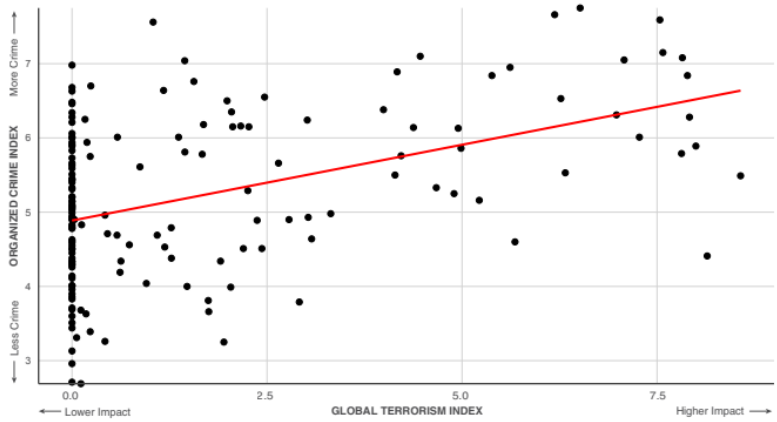

Figure 2.3. Impact of Terrorism vs Organized Crime in 2023. The graph shows that countries with a higher impact of terrorism tend to have higher levels of organized crime [3].

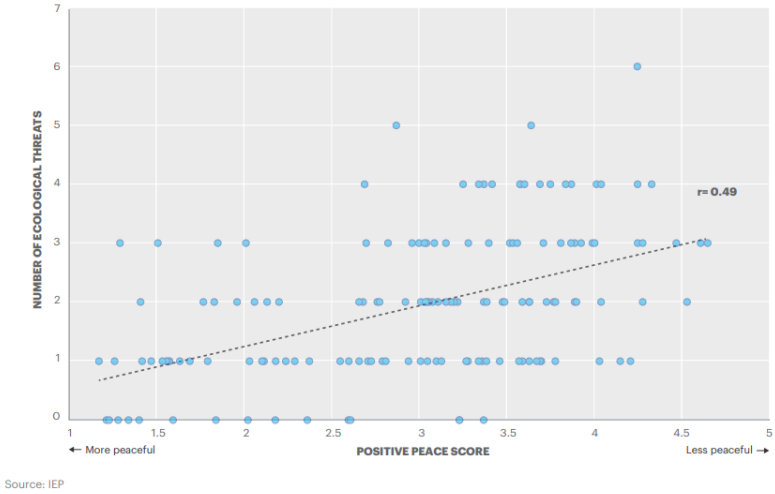

Figure 2.4. Countries with low Positive Peace are exposed to a larger number of ecological threats [21].

2.3.2 Conflicts and Water Scarcity

Water scarcity in the Sahel is one of the major causes of conflicts, although it is complex, deeply entangled in socio-economic issues. [42] highlights that scarcity of water especially in the dry and pre-monsoon seasons is directly proportional to an increase in violence within the Sahel region. By employing indices like the Green Water Scarcity Index (GWS) and the Falkenmark Index, the research shows significant soil moisture deficiency and water stress during critical periods, particularly in areas vulnerable to conflict. This suggests that water shortages driven by population pressures, along with ecological limitations, play an important role in escalating conflicts within the Sahel. Regions that experience water scarcity leading to violent conflicts are usually characterized by population pressures which exacerbate natural drought conditions, leading violent disputes as a result of competition over scarce resources. However, besides the socio-economic aspects, water scarcity cannot adequately explain conflicts without considering such as governance and population increase, which that also contribute in the escalation of water-related conflicts within the region [42]. As further emphasized by [20], water scarcity in the Sahel enhances the danger of migration-related conflict. Another reason this happens is because diminished access to arable land and freshwater heightens competition between farming and herding communities. They have to expand their reach into newer areas because of increased climate pressures, often leading to conflicts over water and land rights. These territorial disputes escalate easily, adversely affecting food security and livelihoods. The connection between water stress and conflict is not limited to Africa or the Sahel region. [21] points that the incidences of water being a cause or a trigger of violence rose by 270% between 2010 and 2018. Moreover, in addition to being a current issue, water scarcity and its consequences are expected to increase. 2.5 illustrates the difference in water demand in the Sahel between 2020 and 2100 [66], highlighting the most affected

areas.

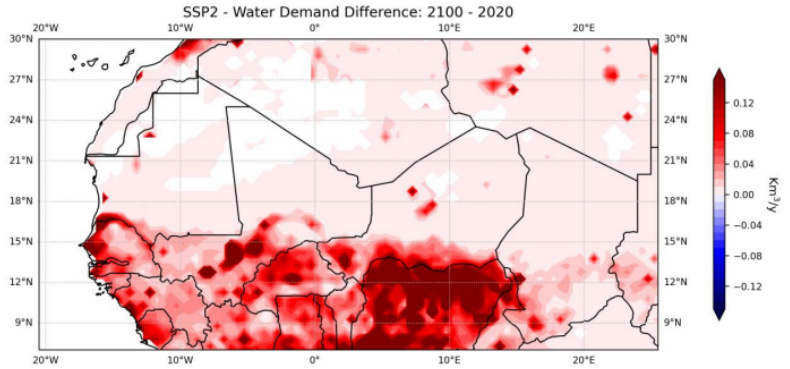

Figure 2.5. Projected difference in water demand in the Sahel between 2020 and 2100, indicating areas most affected by increasing water scarcity [66].

2.3.3 Overpopulation

Overpopulation is putting immense pressure on the region's fragile ecosystems with limited resources. This region currently has one of the highest rates of population increase globally and it is likely to almost double within few decades. Niger, which currently has an estimated population of 26.3 million, is expected to grow in size by 175% up to the year 2050. This fast growth is driven by the very high birth rate and limited access to family planning, leading to strain on infrastructure and public services and, consequently, a lack of educational and economic opportunities, especially among young people. This is well illustrated by the projected trends in population growth, as presented in 2.6. These factors serve to fuel migration and increases the potential for conflict, where competition for resources becomes serious. As noted by a number of studies [1, 62], the convergence of issues like overpopulation, poverty, and resource scarcity have created fertile ground for instability in the region; many youth have lent themselves to extremist groups or engaged in illicit activities due to limited prospects. Addressing the population pressures in the Sahel requires a broad range of interventions, from improving access to education and healthcare to creating economic opportunities, alongside better management of the region's limited resources.

2.3.4 Climate Change and Migration

The Sahel region is especially prone to climate change, with temperature increases at a rate 1.5 times higher than the world average, according to the United Nations [44]. As recently suggested [38, 40], climate-induced migration in the Sahel is primarily driven by factors such as droughts, desertification, and erratic rainfall patterns, which disrupt livelihoods dependent on rain-fed agriculture and pastoralism. Such environmental changes are strong "push" factors that force people to migrate to have a better livelihood. These

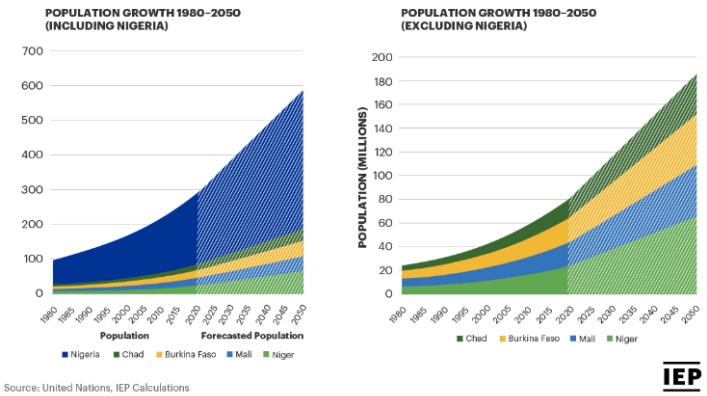

Figure 2.6. Projected population growth in some Sahel countries from 1980 to 2050. The graph shows a significant upward trend [21].

studies also demonstrate that displacement due to climate change is not merely an environmental issue but also one of human rights in terms of access to food, water, and education for the most vulnerable. [38] discusses not only climate change but also a host of socio-political and economic determinants that have built up the intricate patterns of migration in the region. In fact, other important factors of migration include poverty, weak governance, and political instability. Much of the Sahel has very poor infrastructure and a severe shortage of social services, pushing communities to search for better opportunities. The absence of effective governance often implies that the needs of vulnerable populations are not met, thereby creating an environment in which people feel driven to migrate. Moreover, conflicts between various groups of farmers and herders are fanned not only by environmental stress but also by historical tensions and economic disparities. [38] also points out that the situation of violent extremism and terrorist activities in the region further worsen the migration crisis, as people flee these territories to avoid confrontation with violence. This complex interplay of environmental, political, and economic factors in the Sahel highlights that while climate change exacerbates migration and conflict patterns, it is not the sole driver. In fact, the studies suggest that when considering climate-related data alone, predictions of conflict and migration in Africa show little improvement. As such, it is essential to recognize the broader social dynamics at play in the region.

2.3.5 Poverty, Economic and Political Instability

The issues of poverty, economic and political instability are inextricably linked and especially present in the Sahelian region, according to Daniel Eizenga's [18] discussion on the long-term trends of security and development within this region. Most countries in the Sahel have extreme levels of poverty because of limited economic diversification, as well as reliance on agriculture and livestock that makes these countries highly vulnerable to global market fluctuations. This fragility is deepened by the problems of governance

and political instability, where governments are finding it hard to maintain public services and deal with social change. In turn, this political instability mostly heightens the economic challenges by deterring investment and hindering development. Worse, these destabilizing factors have contributed to the rise of violent religious extremism and protracted conflicts in the Sahel, as in many other regions, breaking down social order and impeding development. Together, poverty, economic and political instability form a cycle that puts sustainable development and security in these areas under threat.

Chapter 3

Conflict Early Warning Systems: State of the Art and Ethical Considerations

Early Warning Systems (EWS) represent a critical tool in conflict prevention and management, particularly in regions characterized by complex socio-political dynamics such as the Sahel. These systems combine various data sources and analytical methods to predict and potentially prevent the escalation of conflicts. However, the implementation of such systems raises important ethical considerations that must be carefully addressed. This chapter examines first the current state of EWS in Africa, and subsequently the ethical implications of their deployment.

3.1 African Early Warning Systems

In this section we will explore the state of the art for the EWS examining their functionalities in monitoring potential risks, gathering intelligence, and supporting policymakers in making timely decisions to prevent escalation. Each analyzed system has a different scope, designed to deal with the specific challenges faced in diverse regions of Africa.

This section focuses on the African Union's Continental Early Warning System (CEWS), one of the most important mechanisms to collect data across different regions of Africa to help African Union (AU) decision-makers predict and possibly prevent conflicts. It will also present the ECOWAS ECOWAS Early Warning and Response Network (ECOWARN), which targets conflict-prone areas in West Africa, with emphasis on the Sahel and local indicators of instability.
Next, we will examine the IGAD's Conflict Early Warning and Response Mechanism (CEWARN), which covers East Africa and relies on community-based data to address such specific challenges as resource disputes and cross-border tensions.
Then, we will outline the data-driven model of ViEWS, using machine learning to forecast the probabilities of violent conflicts in Africa, including the Sahel, over a long-term horizon.

Finally, the chapter will present more extensive, worldwide early warning systems, such as the Global Conflict Risk Index (GCRI) and the Water, Peace, and Security (WPS) program, which offer additional context for the examination of conflict risks where environmental and resource challenges converge with violence.

3.1.1 Continental Early Warning System

The CEWS is a prevention initiatives within the African Union. The system, established by the Peace and Security Council, helps the AU staff to anticipate, prevent, and manage conflicts before they escalate. This system was designed to consolidate data from regional and national sources across Africa, creating a centralized framework to monitor potential conflict zones. CEWS employs a range of quantitative and qualitative methods to evaluate indicators spanning political, economic, and environmental domains. These methods aim to produce reliable risk forecasts across the continent.

One of CEWS' key components is its data repository, which collates information from both AU member states and regional organizations, enabling a comprehensive, continent-wide view of emerging threats. However, achieving consistency and standardization across diverse regions remains challenging, as data collection practices and politics will vary by member state. Consequently, CEWS faces obstacles in harmonizing data and ensuring timely responses across regions with different levels of technological and infrastructural development.

Despite these challenges, CEWS has facilitated significant improvements in the coordination of African peacekeeping efforts and strengthened the AU's role as a unifying force in conflict prevention.

3.1.2 Early Warning and Response Network

The ECOWARN is the early warning system specifically targeting areas within West Africa that are vulnerable to conflict like the Sahel. The ECOWARN system operates a network of five regional monitoring centers that provide data collection and assessment related to political, social, and security-related dimensions. The system is uniquely positioned to provide real-time, ground-level intelligence on early indicators of instability, such as electoral violence, civil unrest, and resource-driven tensions. Unlike other early warning systems, ECOWARN incorporates extensive qualitative data gathered from local monitors, which are verified and compiled into risk assessments for use by Economic Community of West African States (ECOWAS) leadership and member states.

However, ECOWARN's operations face some limitations. Their operations are sometimes limited by political constraints within ECOWAS countries, which can affect data transparency and hinder the objectivity of some reports. Additionally, cross-border data sharing is sometimes inconsistent, posing obstacles to comprehensive regional monitoring. As noted by [49] and [2], ECOWAS member states have generally been unable to respond promptly to identified security concerns occurring within their own countries.

This study highlights that early warning systems can only realize their intention if followed by timely interventions; non-compliance in this respect severely lower their impact on the achievement of intended objectives.

3.1.3 IGAD Conflict Early Warning and Response Mechanism

The Intergovernmental Authority on Development (IGAD) established the CEWARN [45] to respond to specific conflict-related issues that are prevalent in East Africa and the Horn of Africa. CEWARN is particularly relevant for countries such as South Sudan and Somalia, where conflicts are usually driven by resource shortages, land, and rivalries between ethnic groups. It collects data on incidents like cattle raiding and other resource-based conflicts that can escalate into wider regional instability. CEWARN has a community-based data collection system, whereby regional contributors convey to the organization specific information on emerging tensions. This approach allows the organization to identify transnational conflicts at a rapid pace since these normally emanate in the pastoral areas of IGAD member states. However, the dependence on local contributors and qualitative assessments creates variabilities in the quality of the data, while the irregular infrastructure across the region may also impede the normalization of data.

3.1.4 ViEWS (Violence & Impacts Early-Warning System)

The Violence Impacts Early-Warning System (ViEWS) is an advanced, award-winning conflict prediction model that forecasts the likelihood of violent conflicts across Africa and the Middle East [25]. It provides monthly probabilistic estimates of armed conflict up to 36 months in advance.

ViEWS is uniquely data-driven and incorporates hundreds of conflict-related indicators, including short-term factors (e.g., recent political events) and long-term structural variables (e.g., institutional strength, demographics, climate impact). The model codifies these diverse datasets into a structured database, organized using Peace Research Institute Oslo Grid (PRIO-GRID) cells and country identifiers to enable sub-national and country-level analysis. By managing input data and transforming features into thematic sets, ViEWS facilitates forecasting and supports user-specified queries on conflict drivers and their predicted impacts.

The forecasts provided by ViEWS are visually interpretable through tools that illustrate relationships between indicators and their joint effects on potential fatalities. These tools aid conflict analysts and policymakers in making data-driven decisions for conflict mitigation. Figure 3.1 provides an example of ViEWS predictions for state-based violence fatalities in February 2024, illustrating forecasted trends and regional risk levels.

ViEWS stands out due to its extensive use of machine learning algorithms, which leverage historical data and real-time updates to continuously improve prediction accuracy. With sophisticated ensemble models and a 36-month forecasting horizon, ViEWS has

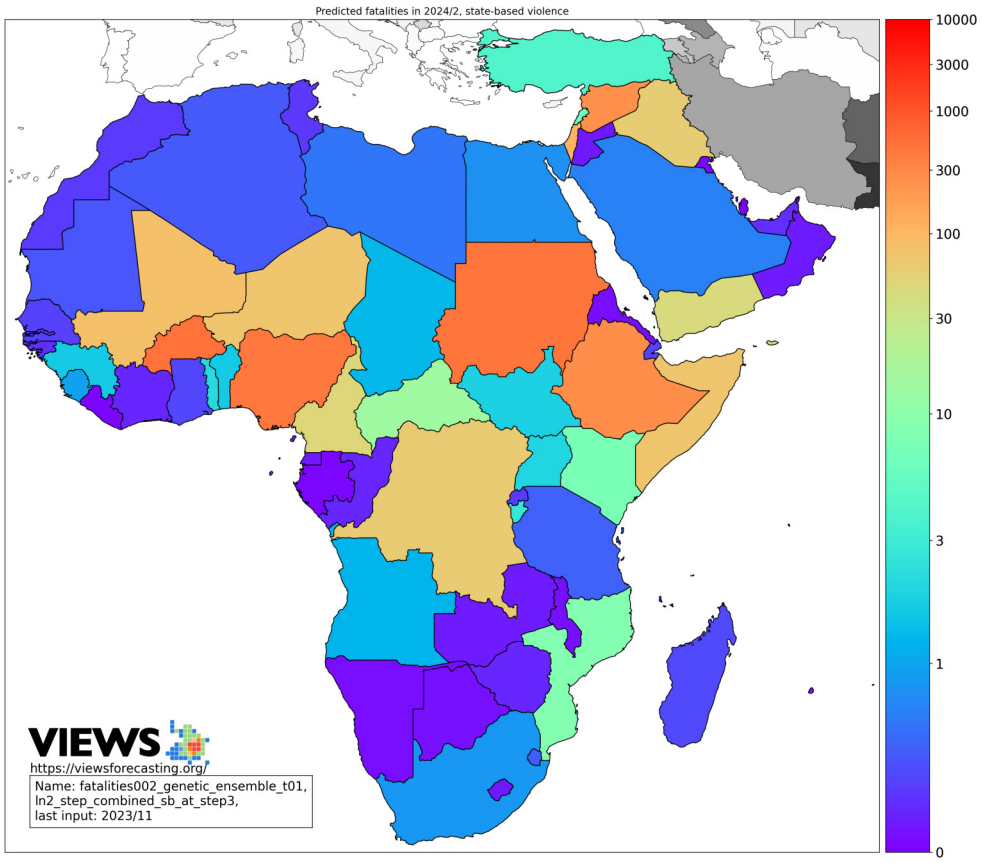

Figure 3.1. Example of ViEWS system forecast for state-based violence fatalities in February 2024 [63].

proven highly effective in assessing risk across multiple levels of violence. Its integration of dynamic, localized data points provides unique insights into emerging conflicts in Africa, including the Sahel, facilitating proactive interventions by governmental and humanitarian stakeholders.

3.1.5 Other Prominent Early Warning Systems: Global Conflict Risk Index and Water, Peace, and Security (WPS)

Other global early warning systems, such as the GCRI and the WPS initiative, also provide valuable insights relevant to the African context. GCRI highlights quantitative modeling methodologies aimed at predicting conflicts at a national and subnational level, especially in regions with a history of political instability and violence. The detailed database and logistic regression models allow for the prediction of risks over periods of up to four years, which considerably contributes to the understanding of dynamics in conflicts within African countries.

The WPS initiative, established in 2018, focuses on conflicts directly linked to water scarcity and climate risks. In Africa, as analyzed in 2.3.2, water is one of the major sources of tension, with about 40% of intra-state conflicts being related to natural resources [35]. WPS provides early warnings of potential violence triggered by water shortages through an integrated approach combining hydrological, stakeholder, and conflict analysis. Its use of geographic and climate data makes it particularly relevant for assessing risk in regions like the Sahel, where resource scarcity acts as a significant conflict escalator.
For instance, in Mali, WPS's analysis revealed how reduced water availability in the Inner Niger Delta creates feedback loops that escalate conflicts: fewer resource reduce income, which leads to over-exploitation and further depletion [35]. This case demonstrates how water stress interacts with social, economic, and political factors to influence conflict dynamics in the Sahel region.

3.2 Ethical and Social Considerations

A conflict early warning system in the Sahel region can present various ethical and social challenges that have to be taken into account during the development of the model. While such a system might enhance initiatives aimed at conflict prevention, its implementation requires a highly nuanced approach that balances the potential benefits with the possible risks and unintended consequences.
Table 3.1 shows the main ethical considerations and some possible corresponding mitigation measures across the phases of the system development.

This section will discuss the main ethical considerations, including the elimination of adversarial adaptation, understanding societal and economic implications, addressing representational biases, and establishing robust governance frameworks.

3.2.1 Mitigating Risks and Unintended Consequences

One of the main ethical issues is adversarial adaptation: malicious agents would like to intentionally deceive the prediction model [7]. This threat may be lower for small-scale groups but it becomes a real possibility for state-level organizations with significant resources and motivations to avoid detection. This underlines the need to restrict the complexity and scope of the dataset used to train the model, in line with the principle of data minimization under current regulations like the General Data Protection Regulation (GDPR).
Another relevant factor to explore is the long-term economic and societal implications on the target countries. [8] considers a hypothetical scenario where the early warning system incorrectly predicts an intense conflict or terrorist attack with 12-month lead time. This long lead time could provoke cancellations of investment, events, and other economic activities in the affected country. Also, it can have disastrous effects on a nation's development and freedom, potentially widening the divide between the target country and more peaceful regions. In the long run, it can reinforce biases and instability that the EWS was supposed to suppress in the first place. The designers of the early warning

Phase	**Main Ethical Considerations**	**Mitigation Measures**
Data Collection	- Privacy Violations - Bias in Data Sources - Unintended Surveillance - Lack of Transparency in Data Processing	- Anonymize and minimize data collection - Apply bias detection - Ensure transparency about data use - Clearly document all data processing steps
Model Training	- Algorithmic Bias - Data Minimization vs. Accuracy Trade-off - Accountability in Decision-Making	- Conduct fairness audits - Balance data minimization with accuracy needs - Define roles and responsibilities for decision-making based on predictions
Model Deployment	- Misuse of Predictions - Economic and Social Impact of Errors - Feedback Loops Amplifying Conflict	- Establish usage guidelines and access control - Regularly assess impact - Periodically review model predictions to prevent feedback loops
Model Monitoring	- Bias Shift Over Time - Adversarial Manipulation - Long-term Social Impact - Accountability in Long-term Impacts	- Continuously recalibrate models for fairness - Implement robust anomaly detection - Conduct long-term impact assessments - Establish accountability for monitoring the system's ongoing impact

Table 3.1. Possible ethical risks across this project, along with their mitigation measures

system should adopt different measures to minimize the risks discussed, as detailed across the phases in 3.1.

These considerations range from the technical aspect of minimizing dataset complexity in order to reduce potential adversarial manipulation, to procedural elements involving rigorous impact assessments for understanding long-term implications. Furthermore, strong privacy and data rights protection must be implemented, along with clear oversight mechanisms and guidelines to prevent misuse.

3.2.2 Representational Bias and Algorithmic Fairness

The development of these predictive models must consider issues of representation and algorithmic bias. The training data used to develop these models could inadvertently reflect historical inequities and power dynamics within the region. For instance, the available data sources like news reports or social media may overrepresent certain geographic areas, ethnicities, or genders, and underrepresent others. In such a case, predictive models may only perpetuate, or even worsen, the marginalization of certain groups in the Sahel.

If such models are not designed with considerations for fairness and inclusivity, they run the risk of creating biased forecasts that may overwhelmingly affect some communities. This may make the already existing inequalities even worse [14].

3.2.3 Governance

The early warning system will also require strong governance frameworks and oversight mechanisms to ensure that it is used responsibly and equitably. This could include the establishment of multi-stakeholder advisory boards, clear data-use policies, and independent auditing processes. Decision-making authority and possible misuse by powerful actors should be considered with care [39].

3.2.4 Ethical Considerations of Online News and Social Media Data

The use of publicly available data from online news sources and social media platforms, such as Twitter, raises major ethical considerations that must be taken into account.
These sources can carry inherent biases and privacy risks [12]. Online news reporting may reflect the biases of the media organizations or government agencies that produce it, consequently affecting the accuracy of predictions by the model. The media coverage then might tend to disproportionately highlight some areas or demographic groups influenced by geopolitical priorities or cultural views, therefore introducing the tendency of the model to overrepresent particular incidents and neglect others. Moreover, relying on online news sources may reinforce existing narratives, thereby creating feedback loops that maintain stereotypes about the Sahel region.
In a similar way, the use of Twitter data for prediction purposes introduces ethical concerns. While social media can provide immediate insights into regional sentiments and emerging tensions, their use brings up questions about user privacy since posts, while publicly available, had not been intended for predictive surveillance [51]. Incorporating Twitter data into the forecasting of conflicts may inadvertently infringe on the rights of people or groups to express themselves freely [48]. Moreover, malicious actors could manipulate social media platforms with false information in order to change the predictions made by such models, raising ethical questions about accountability and data integrity [65].
In order to mitigate these issues, the principle of data minimization advocated by the GDPR should guide the collection and the use of online news and social media data [46]. By limiting the scope of data and concentrating solely on the most critical information, researchers and policymakers can contribute to the assurance that the implementation

of the early warning system respects individual rights and diminishes the potential for misuse. Also, transparency regarding the sources of data and frequent auditing of the datasets and algorithms will be necessary to ensure that ethical principles are upheld [36].

Chapter 4

Datasets and Data Collection

In this chapter, we provide a detailed overview of the datasets analyzed and used in this work, their sources and the most relevant features, in order to improve the fatalities prediction in the Sahel region.

4.1 GDELT Dataset Overview

Global Database of Events, Language, and Tone (GDELT) is an open-source dataset that contains events, interactions, and actions across media channels around the world. It monitors in real-time broadcast, print, and web news capturing various event attributes, including geographical location, event type, actors involved, and tone.
This is helpful for understanding the dynamics of conflict, since it gives real-time insight into the social, political, and military events happening in these regions. The database is updated every 15 minutes, which allows researchers ability to track global events as they happen.
Figure 4.1 plots the trend in the number of online news articles in the GDELT database for Sahelian countries. The y-axis is scaled to display values in millions (10^6). Nigeria has been removed from the graph to improve visibility; however, it follows a trend in date and article numbers similar to the other countries. It is important to notice that the number of articles increased sharply beginning in 2008 and peaked in 2015, aligning with the launch of version 2 of the GDELT dataset. In the following years there has been a significant drop in the number of articles. This decline is not limited to the Sahel countries, as similar trends have been observed in other regions and for other data types as well; it is a characteristic feature of GDELT data. For our purposes this decrease does not necessarily pose a problem, but it has to be kept in mind when carrying out subsequent analyses.

The GDELT data was accessed and downloaded via queries on Google BigQuery, filtered to only include records from 1998 onward in order to be compatible with the other datasets. Only events occurring within the Sahel region were selected using the Action-Geo_CountryCode field with an emphasis on events that could signal conflict. To further refine this selection we filtered the GoldsteinScale variable to only include events with

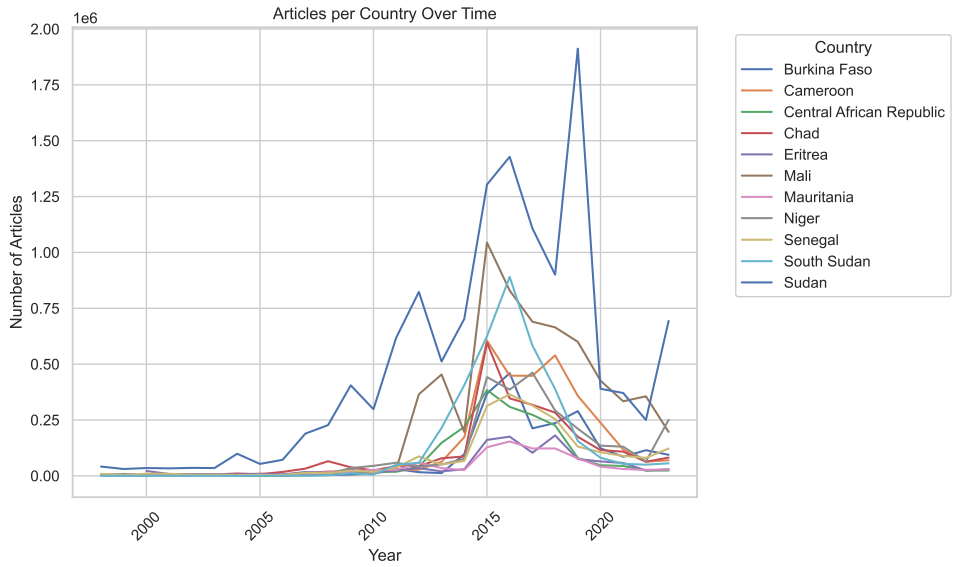

Figure 4.1. Trend in the number of online news articles in the GDELT database for Sahelian countries. It shows a significant increase starting in 2008, peaking in 2015, and then a gradual decline. This reduction is repeated in other regions or datasets and indicates a systemic change in the GDELT dataset, not dependent on Sahelian countries.

scores lower than 0, as these can represent conflictual events [56]. The Goldstein scale is an established metric for the conflictual or cooperative nature of international interactions, though it has limitations, including treatment of the intensity of events. The Goldstein score applies only to the action category of an event, such as fight or trade, meaning that it excludes other dimensions, including variations in casualties, actors involved, or temporal factors.
This results in different events carrying the same label of "fight" receiving the same Goldstein values despite a possible large discrepancy in severity or impact [56].

Fig.4.2 shows how the average Goldstein Score evolved over time, considering only negative values to capture conflict articles rather than peaceful interactions; the data shows different fluctuation through countries and dates. The initial South Sudan's downward trend is not statistically significant due to sparse media coverage between 1998 and 2005.

4.1.1 Most relevant indicators from the GDELT Dataset

The following indicators from GDELT have been selected for our analysis since they can capture key aspects of event tracking. Each of them provides specific information about the news events and together can offer a detailed view of the events and their context.

- **SQLDATE**: this includes the date of the event in the format 'YYYYMMDD'. Useful in tracking the chronology of events and analyzing trends over time.

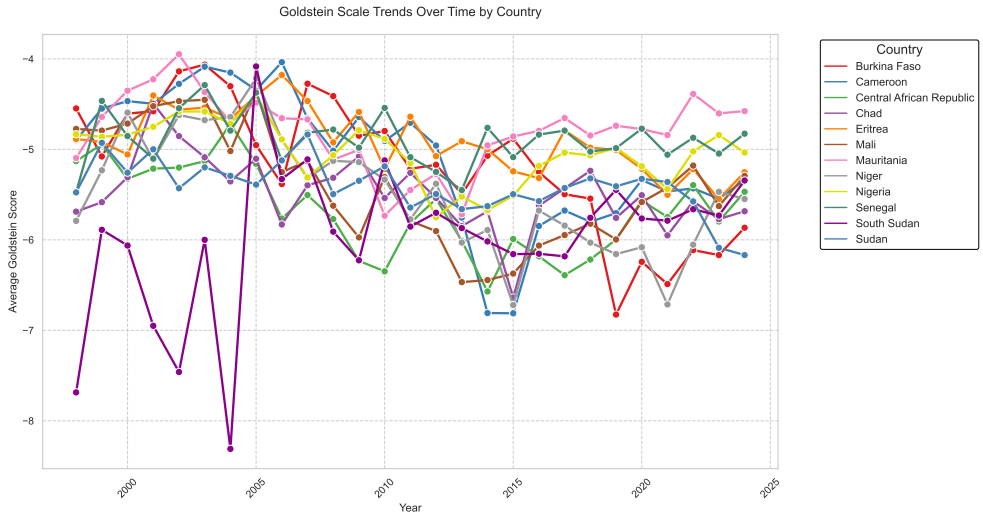

Figure 4.2. Temporal evolution of the Goldstein scale scores across 12 African countries.

- **ActionGeo_CountryCode**: this variable records the country in which the event took place using ISO-2 country codes. It localizes the events to focus specifically on Sahelian countries.
- **NumSources**: this is the number of independent sources that have reported the event. The more sources, the higher the reliability and possible importance of the event. It also helps in rating the media coverage and attention an event receives.
- **NumArticles**: this is the number of news articles that have been used to create the event record. The number of articles can be a proxy for how important or public an event was, particularly in conflict contexts.
- **EventCode**: GDELT uses a coding scheme based on the Conflict and Mediation Event Observations (CAMEO) system. This variable specifies the type of event that occurred, such as protest, violent actions, or diplomatic engagement. This allows for distinguishing conflict events from non-conflict events.
- **QuadClass**: quadClass is divides events into four general classes: verbal cooperation, material cooperation, verbal conflict, and material conflict. It enables examination of the nature of events to see if an event is likely to further conflict.
- **GoldsteinScale**: this variable is a measure of the potential impact of an event on global stability, ranging from highly conflictual to highly cooperative. It helps quantify the intensity and potential impact of events on conflict escalation or resolution.
- **isRootEvent**: a binary variable indicating whether the event is a root event or something that occurs as part of a chain of events. In particular, root events are

the initial action that may spur other events that follow and thus important to understand where conflicts begin.

- **AvgTone**: this variable calculates the average sentiment of the various news about the event, going from -100 to +100. Specifically in our context, a more negative tone is more likely to mirror violence and unrest, whereas a positive tone may reflect cooperation or peaceful resolutions.

- **SourceURL**: URL of the original news report, providing access to the source article text when available and free from restrictions.

4.2 ACLED Dataset

The Armed Conflict Location and Event Data Project (ACLED) [4] is one of the most frequently cited comprehensive datasets, providing real-time data on political violence, protests, and associated events worldwide. ACLED codes many forms of conflict, including armed clashes, violence against civilians, terrorism, riots, and non-violent protests; it captures events at the most granular level possible, often down to the individual incident. Each event is coded with specific attributes, including the date and location (down to the district level), type of event, actors involved, and, in particular, the number of fatalities. The ACLED data is collected from a broad range of sources, including news reports, local media, international organizations, and non-governmental organizations.
Its rigorous methodology ensures that large-scale conflicts are recorded, but so are the smaller, often less visible incidents, providing a dataset invaluable for the analysis of conflict patterns, the spread of violence, and its impact on local populations. Most datasets are updated on a weekly basis, letting one track, in almost real time, the trends of conflicts.

For these characteristics, ACLED has been an essential resource in our analysis. ACLED data can be used to identify patterns and signals that precede violent events, such as increased protests, troop movements, or clashes between armed groups. Moreover, the event-level granularity in the dataset allows for analyzing localized trends, which are crucial for crafting interventions that are tailored to prevent further violence and deaths.

Figure 4.3 displays the trend in fatalities by country. As shown, not all countries have large numbers of fatalities; in particular, Eritrea, Mauritania, and Senegal never reach the level of 500 fatalities in a year; Chad and Niger also only occasionally reach the level of 1,000 fatalities. This is a piece of important information for the model development since the number of fatalities will be a predominant determinant in indicating possible conflict presence. It must be specified that data regarding the Ethiopia-Eritrea War (May 6, 1998 – June 18, 2000), which is estimated to have resulted in between 70,000 and 100,000 deaths, is not included in this graph in order to improve the visualization.

The key features selected for our analysis are:

- **Event_Date**: date in which the event occurred, in the form Year-Month-Day.

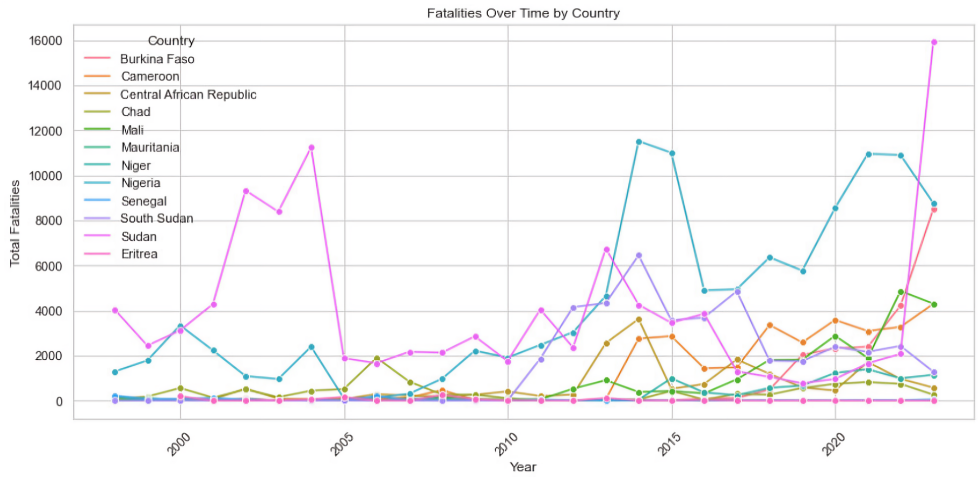

Figure 4.3. Fatality trend by country.

- **Disorder_type**: categorizes the type of disorder, helping distinguish between different forms of conflict.
- **Event_Type**: provides a more detailed classification of events.
- **Sub_Event_Type**: further specifies the nature of events, offering the finest level of event classification.
- **Country**: expresses where the event took place.
- **Fatalities**: number of reported fatalities resulting from an event.

4.3 UNDP Dataset

The United Nations Development Programme (UNDP) [60] is the organization that acts in various sectors to fight against poverty, reduce inequalities, and ensure resilience especially in developing countries. One of its main initiatives is the Human Development Reports (HDR), which serve as an informative resource to monitor the state of human development. The HDR produces a number of composite indices, such as the Human Development Index (HDI) [50], to give a broad view of human development progress across nations.

4.3.1 The Human Development Index

The Human Development Index [50] is a measure that captures average achievements in three fundamental dimensions of human development: health, education, and standard of living.

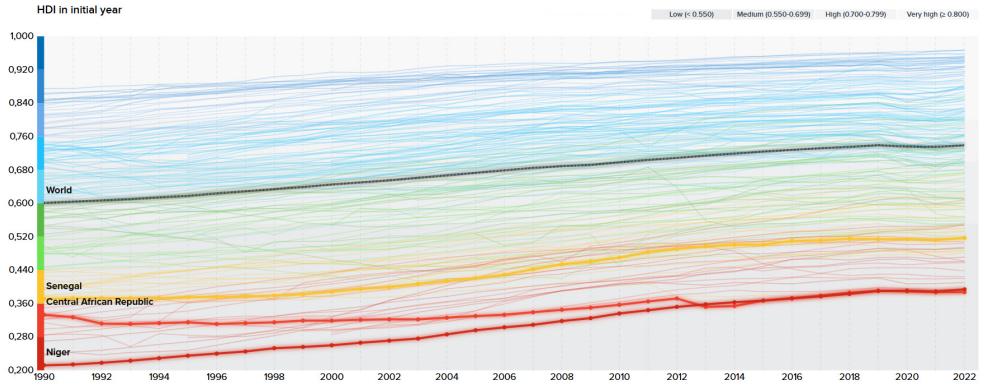

Figure 4.4. HDI trends [59] for three Sahelian countries: Niger, Central African Republic, and Senegal, compared with the global average.

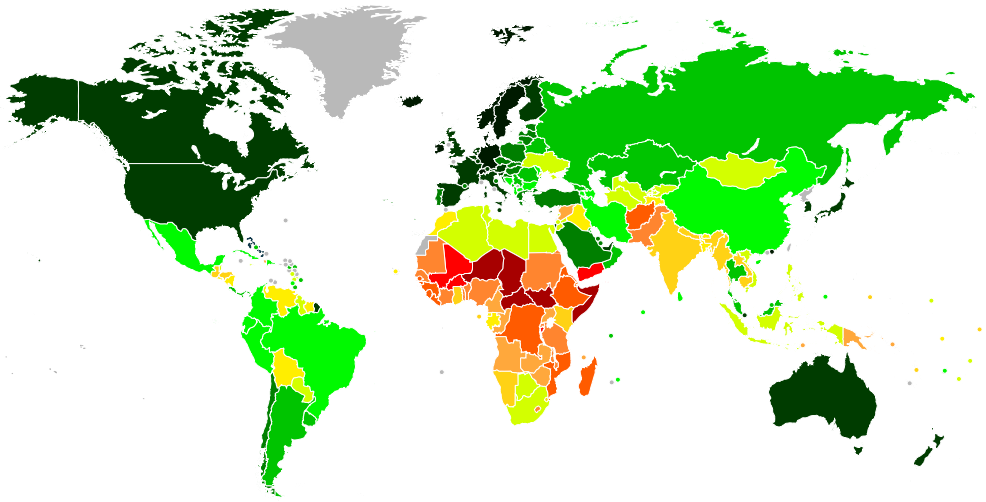

Figure 4.5. Global distribution of the Human Development Index [61]. The map highlights clear disparities, with Sub-Saharan Africa (in red and dark orange) having the lowest HDI values globally.

- **Health Dimension**: this is measured using life expectancy at birth as the proxy to a healthy and long life.
- **Education Dimension**: it is measured using two sub-components: mean years of schooling for adults aged 25 years and older and expected years of schooling for children of school entrance age.
- **Standard of Living Dimension**: this is measured as gross national income (GNI) per head, with the income figures transformed on a logarithmic scale to capture the

declining marginal utility of income as it rises.

The HDI is calculated by taking the geometric mean of normalized indices for each of these three dimensions. The index provides a balanced view of human development and moves beyond economic wealth alone, including social and educational factors.

The relative position of three Sahel countries, namely Niger, Central African Republic and Senegal, is depicted in figure 4.4, comparing with the global average. Generally, the index tends upwards over time, but such progress may be challenged by political instability, economic challenges, social disparities, and corruption. Thus, for instance, the Central African Republic shows minimal HDI improvement over 30 years, as a result of these complex problems. The Sahelian countries consistently show some of the lowest HDI scores worldwide, as shown by the most recent HDI values from 2022 in Figure 4.5.

While very useful, it is important to recognize its limitations. The HDI does not account for inequality, poverty, or human security, factors that may also influence the socio-political dynamics in the Sahel. However, it remains one of the important explanatory variables for understanding and predicting trends of human development despite some of its weaknesses across countries and over time.

4.4 World Bank Dataset Overview

The World Bank Development Indicators dataset [67] is a wide-ranging compilation of development data from national and international agencies. It contains statistical indicators for monitoring global development, including data on population, economic growth, governance, education, health, infrastructure, and environmental sustainability. Regular updates to this dataset are made available for more than 200 economies, making it an important resource for monitoring and identifying trends in development.

4.4.1 Selected Indicators from the World Bank Dataset

For this project, we selected indicators that reflect economic stability, governance, demographic pressures, and environmental stress in the Sahel region to support conflict prediction.

- **Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)**: this indicator shows the stress on water resources, which, as already analyzed in 2.3.2, is especially meaningful in arid areas like the Sahel, where water scarcity can feed conflict.
- **Access to electricity (% of population)**: this measure reflects the availability of electric power to the general population.
- **GDP growth (annual %)**: the economic growth rates provide an insight about the state of the economy. As introduced in 2.3.5, where economic growth is stagnant or negative, the chance of social unrest and conflict becomes greater.

- **Military expenditure (% of GDP)**: measure of the share of national resources allocated to the army. Higher military spending can be an indicator of future conflict.

- **Broad money (% of GDP)**: this indicator measures the amount of money which circulates in the economy. A strong financial system is an indicator of economic stability, which might reduce the conflict likelihood.

- **Population density (people per sq. km of land area)**: high population density can result in inefficient resource utilization, raising the chances of a conflict.

- **Population growth (annual %)**: similar to the previous indicator, high population growth can put pressure on the already limited resources, infrastructure, and services, increasing the potential for conflict.

- **Political Stability and Absence of Violence/Terrorism: Estimate**: indicator of governance which assesses the likelihood of political instability and/or politically motivated violence, including forms of terrorism. Lower scores are associated with a higher risk of conflict.

- **Rule of Law, Estimate**: rule of law measures the extent of confidence in societal rules, including the quality of contract enforcement, property rights, the police, and the courts, as well as the probability of crime and violence.

- **Control of Corruption, Estimate**: this measure assesses people's perceptions of the degree to which government officials and politicians abuse their power for personal gain or corruption.

- **Government Effectiveness, Estimate**: the government's effectiveness is an estimate based on several factors such as the quality of public services, civil service, policy formulation, and the credibility of the government's commitment to policies.

- **Battle-related deaths (number of people)**: this indicator captures the impact of armed conflict on civilian populations.

- **Intentional homicides (per 100,000 people)**: a measure of the level of internal violence, can be indicative of underlying social instability.

- **Internally displaced persons, total displaced by conflict and violence (number of people)**: number of people displaced within their country due to conflict and violence, representing a measure of the human impact of ongoing conflicts.

These indicators provide a complex, multi-dimensional view of the socio-economic and political landscape across the Sahel region. Through the examination of trends in these variables, we can enhance our understanding of conflict drivers.

4.5 Twitter (X)

In order to improve the model with real-time data, we considered twitter data as a complementary source. Social media platforms have emerged as valuable sources for conflict detection and monitoring.
Twitter, in particular, has been widely used in conflict studies due to its public nature and real-time information flow [15, 22]. Twitter can be used for academic purposes, as shown by Twitter's Developer Terms of Service, but the new API policies appear to be increasingly restrictive after Elon Musk's acquisition. In particular, they provide limited search with reduced coverage compared to the actual platform activity.

To maintain data quality within these constraints, we focused our analysis on five countries that are well representative of the Sahel region: Sudan and Cameroon which have high fatality rates, and Chad, Central African Republic, and South Sudan with medium and low fatality rates with respect to the whole area.

To retrieve meaningful tweets for our context and maintain a reasonable dataset size, we constructed the query with "conflict + country name". The query returns not only exact matches but also contextually relevant content. Thanks to Twitter's relevance algorithm, it captures tweets related to conflict events as peace negotiations, humanitarian situations, and social tensions. The considered temporal coverage is from 2010 to 2021, with data after 2021 reserved for testing purposes.

Due to the mentioned API limitations like rate limits and coverage limitations, we implemented multiple collection iterations when the initial scraping didn't yield sufficient data points.

Chapter 5

Data Preparation and Analysis Pipeline

The following chapter reports the analytical steps through which our collected indicators, previously presented in Chapter 4, can be converted into the final features considered by the model. At time X, the model incorporates both current features and various lagged features, as we will detail in Section 5.4. In terms of general characteristics, the model processes 101 features, along with country and date identifiers. An analysis of temporal granularity revealed that weekly aggregation provides better results, yielding 15,724 observations from early 1998 through 2023. Moreover, it can be noted that the most informative data points are concentrated from 2010 and beyond, and they coincide with more availability of comprehensive GDELT news coverage. This weekly aggregation can capture rapidly changing trends, particularly in news-based indicators, while maintaining good generalization capability. Furthermore, this approach significantly reduces computational complexity compared to a daily model, which would require processing seven times the current number of rows but fails in generalization.

5.1 Conflict Events and Fatalities Analysis

5.1.1 Fatalities Considerations

The fatalities can be considered our primary predictor, and for this reason it is necessary to pay special attention to it. The model considers multiple prediction horizons and our target variable is given by the sum of fatalities from the last observable point to the prediction endpoint. As illustrated in Section 4.2, the number of fatalities is unbounded, hence they need an appropriate preprocessing to manage extreme values. Although at first the logarithmic transformation was considered to constrain the range, it did not yield satisfactory results. Instead, more effective results were achieved by truncating fatalities at country-specific thresholds. This truncation approach was implemented by analyzing historical fatality peaks for each country and setting the threshold at the 95-th percentile of the distribution.

This method effectively manages outliers, while preserving the meaningful variation in the data. Indeed, since extremely high fatality counts are often rare, these extreme events could potentially distort the model's predictive capabilities. The truncation is implemented through a maximum threshold computed separately for each country, ensuring that the model remains sensitive to the typical conflict patterns while being robust against extreme outliers.

In Figure 5.1 the heatmap illustrates the monthly correlation of fatalities between countries in the region. Certain countries exhibit moderate to high correlations, such as Burkina Faso and Mali, which suggests a shared pattern in the fluctuations of fatalities, potentially driven by regional dynamics or interconnected conflicts. For example, their strong correlation reflects not only their geographic proximity but also the influence of transnational insurgent groups, such as Al-Qaeda in the Islamic Maghreb, which operate across their borders.
Similarly, correlations between Niger and its neighbors, including Nigeria and Burkina Faso, highlight potential spillover effects of violence across borders or shared conflict drivers in the region. The moderate correlation between Niger and Nigeria could generate from shared regional pressures, such as the impact of Boko Haram in the Lake Chad basin, which affects multiple countries simultaneously. This correlation may also indicate the influence of joint security initiatives or economic interdependence, which can lead to synchronized trends during periods of conflict.

In contrast, certain country pairs display low correlations, reflecting more distinct or localized patterns of violence, For instance, Sudan and South Sudan, despite being neighbor countries, show a low correlation. This is not surprising, as the reasons for this divergence can be various. South Sudan gained independence from Sudan in 2011, and since then, the two countries have experienced distinct conflict trajectories. South Sudan has been haunted by internal civil conflicts driven by power struggles between political factions, while Sudan is currently dealing with its own severe internal conflict between the Sudanese Armed Forces and the Rapid Support Forces, which started in 2023. Also, demographic differences contribute to distinct conflict dynamics. In fact, Sudan has an Arab and Muslim population, while South Sudan has an African Christian and animist population. Economic disparities further amplify this division, with South Sudan retaining the majority of oil reserves from Sudan, creating divergent economic challenges and development paths.

5.1.2 Disorder Types

As mentioned in Section 4.2, the ACLED dataset provides information about disorder events occurring in the Sahel region. Excluding rare events, there are three main categories of disorder events:

- **Political Violence**: violent events with political motives, as battles, violence against civilians or explosions. This often involves conflicts between state forces and non-state armed groups, inter-communal violence, and terrorist activities. These

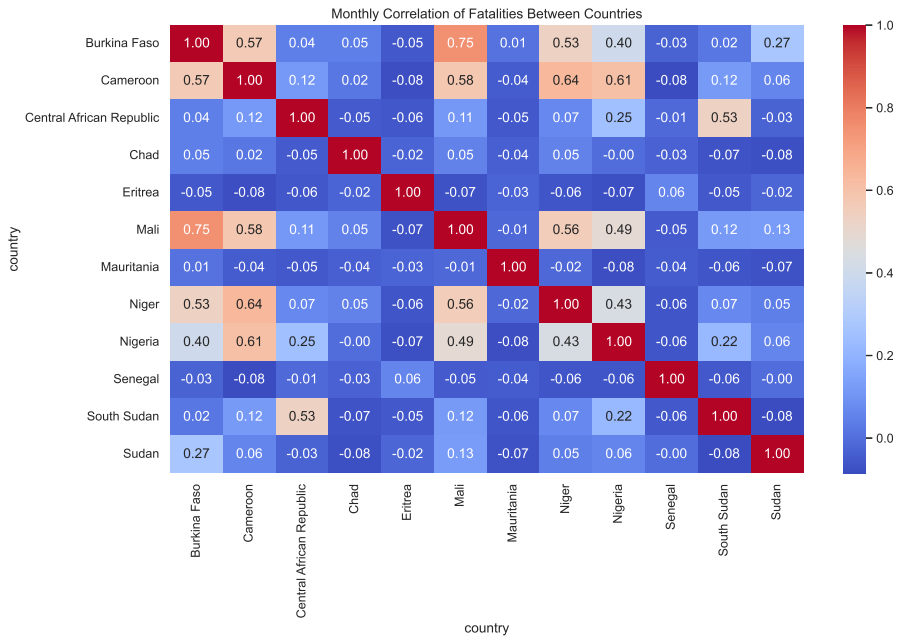

Figure 5.1. Heatmap of monthly correlation of fatalities between countries in the Sahel region.

events typically represent the worst form of disorder.

- **Demonstrations**: protests and riots, collective actions by civilians. While demonstrations can be peaceful, they may escalate into violence.
- **Strategic Developments**: significant events that may not involve direct violence but are crucial for understanding conflict dynamics. Some example include peace agreements, force deployments, or changes in group leadership. They often signal important shifts in conflict patterns and can be precursors to future violence or peace.

These disorder types were encoded using one-hot encoding, transforming the categorical variables into binary features. This type of encoding enables the quantitative analysis of event patterns while preserving the distinct nature of each disorder type.

5.2 Indicators Processing and Analysis

The WDI indicators considered in this study are listed in Section 4.4. Although some indicators were available from 1960, to better align with other datasets, we limited our analysis to data from 1998 onwards. From the 1,488 indicators available in the dataset, we selected 19 of them that could give comprehensive information about each country without overloading our predictive model.

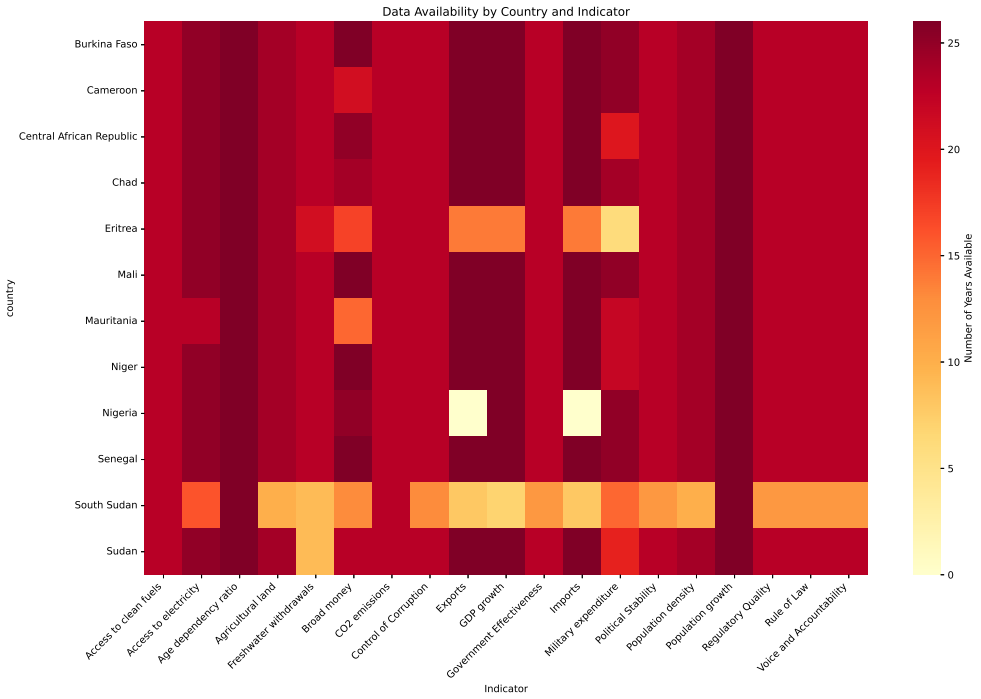

Figure 5.2. Data availability across countries and indicators. Darker red indicates more years of available data.

All nations were generally well represented by these indicators over the years, with few exceptions. Considering the 26-year period from 1998 to 2023, we included indicators for countries that had at least 18 years of data available. This threshold of 70% data availability was chosen to ensure statistical robustness while maintaining enough data points for meaningful temporal analysis.

5.2.1 Indicators Availability

Figure 5.2 shows the data availability across countries and indicators, highlighting particular gaps in South Sudan's data and sporadic missing values in other countries, especially for indicators like freshwater withdrawals and CO2 emissions.

For the remaining missing values within our selected timeframe, we applied linear interpolation to maintain data continuity. This method was chosen over simple forward or backward filling as it better preserves trends in the data. For a missing value y at time t between two known values y_1 and y_2 at times t_1 and t_2 respectively, the interpolated value is calculated as:

$$y = y_1 + (y_2 - y_1) * \frac{t - t_1}{t_2 - t_1} \tag{5.1}$$

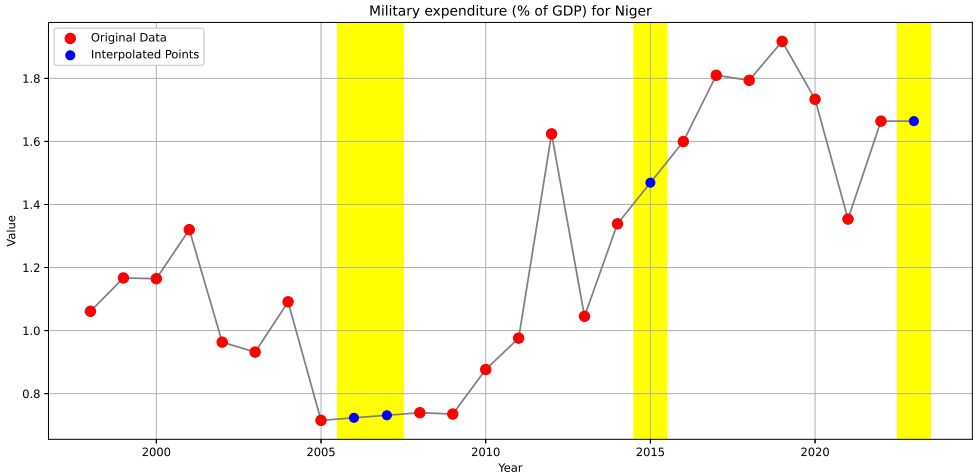

Figure 5.3. Example of linear interpolation applied to Niger's military expenditure data. Red points represent original data, blue points show interpolated values, and yellow bands highlight periods where interpolation was applied.

This approach assumes a linear relationship between consecutive known values. Figure 5.3 illustrates this interpolation process using Niger's military expenditure data as an example. The visualization shows original data points (in red), interpolated values (in blue), and highlights the periods where interpolation was applied (yellow bands). This example demonstrates how linear interpolation preserves the underlying trends while filling gaps in the time series, providing a continuous and plausible trajectory of the indicator's evolution.

Particular attention should be given to the HDI indicator, since it is a composite indicator that embodies key dimensions of human development. The graph in Fig. 5.4 presents its the temporal evolution values across all Sahelian countries in our study from 1990 to 2022. There can be noticed significant variations in both the starting points and trajectories across the region.

5.3 Feature Engineering and Selection

5.3.1 GDELT Features Construction

The GDELT Database provides daily news-based indicators of conflict-related events. Its real-time nature and daily updates make it particularly valuable for our context, as it enables rapid detection of possible conflict patterns. By elaborating the dataset features we retrieved the following metrics:

- **Event Intensity Metrics**
  - *Goldstein_weighted*: A weighted average of the Goldstein scale values, which

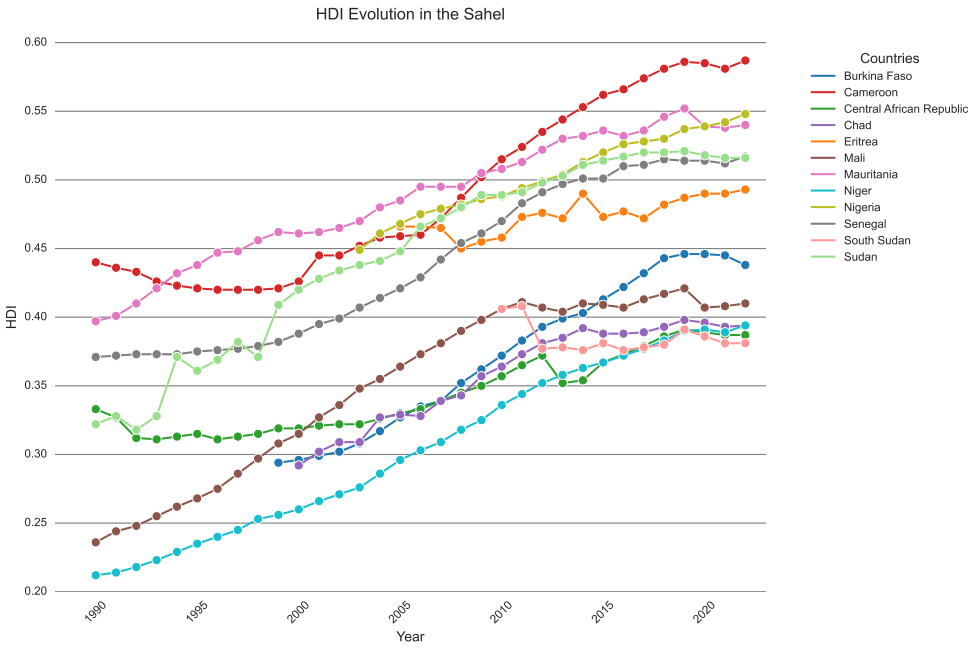

Figure 5.4. Evolution of Human Development Index (HDI) in Sahelian countries from 1990 to 2022.

range from -10 to 0. The weights are based on the number of articles, giving more importance to events with higher media coverage:

$$Goldstein\_weighted_w = \frac{\sum_{d \in w}(Goldstein_d \times NumArticles_d)}{\sum_{d \in w} NumArticles_d} \tag{5.2}$$

where w represents the week and d represents the daily values.

– *Goldstein_trend*: Captures the directional change in event intensity:

$$Goldstein\_trend_w = Goldstein\_weighted_w - Goldstein\_weighted_{w-1} \tag{5.3}$$

- **Media Attention Indicators**

  – *NumArticles_normalized*: Article counts normalized by a 6-month moving window to account for temporal biases in media coverage:

$$NumArticles\_normalized_w = \frac{NumArticles\_total_w}{\text{rolling\_mean}_{26w}(NumArticles\_total)} \tag{5.4}$$

  – *media_pressure*: Combines normalized media attention with event intensity:

$$media\_pressure = NumArticles\_normalized \times |Goldstein\_weighted| \tag{5.5}$$

  - *media_acceleration*: Second-order changes in media attention, capturing rapid shifts in coverage

- **Temporal Pattern Features**
  - *sustained_attention*: Consecutive weeks of above-average media attention
  - *media_volatility*: Standard deviation of media attention over a 4-week window
  - *goldstein_stability*: Rolling standard deviation of event intensity

5.3.2 Processing Pipeline

The feature engineering process involved several key steps to transform raw GDELT data into meaningful predictors:

1. **Temporal Aggregation**: Daily events were aggregated into weekly intervals to reduce noise while maintaining temporal resolution. For event metrics, weighted averages (Goldstein scale) were used to preserve the significance of heavily covered events.

2. **Rolling Statistics**: We implemented 4-week rolling windows to capture recent patterns:
$$X\_rolling\_mean_w = \frac{1}{4}\sum_{i=0}^{3} X_{w-i} \tag{5.6}$$
where X represents various metrics (fatalities, media coverage, etc.)

3. **Media Coverage Normalization**: A 26-week rolling window was used to normalize article counts in order to address long-term trends in digital media coverage, country-specific variations in media attention, and temporal biases in GDELT coverage.

4. **Pattern Detection**: Sustained patterns were identified using consecutive occurrence counting. To better identify patterns, volatility metrics were computed using rolling standard deviations, and crisis intensity was calculated by combining media pressure with fatality counts.

5.3.3 Feature Refinement

Initial Feature Cleaning

The initial set of features of GDELT was refined through a two-step selection process. In the first step of the selection process, we removed features that were considered less informative for conflict prediction. From the media-related features we removed *tone_weighted* and *tone_trend* since their informations can be found in the other indicators. *NumArticles_total* was replaced by normalized versions to account for temporal biases, while *NumArticles_trend* has been considered redundant since the information could be captured in other media dynamics features. As regards the conflict-related features, we removed *fatalities_acceleration* since it showed high volatility and less stability compared

to rolling statistics, and *quarter* which led to redundant temporal information already captured by *month*.

Multicollinearity Reduction and Final Features

As regards the second step of the selection process, we addressed multicollinearity by removing highly correlated features. Among the rolling fatalities statistics (*rolling_sum*, *rolling_mean*, *rolling_max*), we retained only *fatalities_rollling_mean* as it provides the most balanced view of conflict intensity. *NumArticles_normalized* was removed since *media_pressure* already captured all the important information. Finally, *media_acceleration* was eliminated due to its similarity with other trend metrics.

The final feature set consists of:

- Event intensity metrics: *Goldstein_weighted*, *Goldstein_trend*, *Goldstein_rolling_mean*, *goldstein_stability*
- Media coverage indicators: *media_pressure*, *media_volatility*
- Conflict patterns: *fatalities_rolling_mean*, *crisis_intensity*
- Temporal patterns: *sustained_attention*, *quiet_period*, *month*

This refined set of features provides a comprehensive view of conflict dynamics while maintaining computational efficiency and avoiding redundancy in the predictive model.

5.3.4 Dimensionality Reduction

Correlation Analysis

To manage the high dimensionality of the socio-economic indicators while preserving their informative value, we implemented a mixed approach combining dimensionality reduction and feature selection. First, we analyzed the correlation patterns within groups of related indicators:

- Development indicators (clean fuels access, electricity access) showed a high correlation (explained variance 0.83)
- Governance metrics (control of corruption, government effectiveness, regulatory quality, rule of law, voice accountability) displayed strong internal relationships (explained variance 0.78)
- Resource indicators (agricultural land, freshwater agriculture) exhibited high correlation (explained variance 0.82)
- Economic indicators showed weak internal correlations (explained variance 0.36)
- Population metrics displayed moderate correlations (explained variance 0.49)

PCA and Final Features

Based on the previous analysis, we applied Principal Component Analysis (PCA) to create composite features for highly correlated groups. PCA was particularly effective for groups with high explained variance (>0.75), indicating that a single composite feature could capture most of the variability in the original indicators. For each group, we standardized the features and extracted the first principal component, which represents the linear combination that explains the maximum variance in the data.
Specifically, we created composite features for development indicators (*development_composite*), resource indicators (*resource_composite*), and governance indicators (*governance_composite*). The features that we maintained as separate are: economic indicators (*broad_money_gdp*, *exports_gdp*, *imports_gdp*, *gdp_growth*) due to their low internal correlations (explained variance 0.36), suggesting they capture distinct aspects of economic conditions that might influence conflicts differently; population metrics (*age_dependency_ratio*, *population_density*, *population_growth*) since these metrics represents different demographic dynamics that could independently affect conflict patterns although the moderate correlations (explained variance 0.49); political stability (*political_stability*) has been separated from other governance indicators as it showed a lower correlation with them and directly measures conflict-related aspects, making it relevant for our prediction task.
This approach reduced the feature space from 17 to 11 features while preserving the distinct information captured by weakly correlated indicators. The PCA-based composite features effectively summarize groups of correlated indicators, while the separation of weakly correlated features ensures no significant information is lost in the dimensionality reduction process.

5.4 Lag Indicators and Final Indicators

The final set of base indicators consists on:

- Event-based metrics (*goldstein_weighted*, *goldstein_trend*, *goldstein_stability*, *goldstein_rolling_mean*)
- Media attention metrics (*media_pressure*, *media_volatility*, *sustained_attention*, *quiet_period*, *numarticles_rolling_mean*)
- Conflict indicators (*crisis_intensity*, *fatalities*, *demonstrations*, *political_violence*, *strategic_developments*)
- Socio-economic indicators (*age_dependency_ratio*, *broad_money_gdp*, *exports_gdp*, *gdp_growth*, *imports_gdp*)
- Development and governance metrics (*military_expenditure*, *political_stability*, *population_density*, *population_growth*, *development_composite*, *resources_composite*, *governance_composite*, *hdi*)
- Temporal indicators (*month*)

Because of the time series nature of our data, we incorporated information from previous time periods in order to capture temporal patterns and trends. Lag indicators were created specifically for features that exhibit significant temporal variation and where changes over time are particularly informative. The lag structure includes: 1-week lag, 2-week lag, aggregated 3-4 week lag, aggregated 5-8 week lag, aggregated 9-12 week lag, and similar aggregations up to week 16. This hierarchical lag structure allows the model to capture both immediate and longer-term temporal dependencies.

The following features were selected for lag creation:

```
fatalities
political_violence
demonstrations
crisis_intensity
goldstein_weighted
goldstein_trend
media_pressure
media_volatility
sustained_attention
numarticles_rolling_mean
political_stability
strategic_developments
```

In total, the model incorporates 101 features, plus country and date identifiers.

Chapter 6

Experiments and Results

In this chapter, we will explore the various components that led to our final results. In the first sections we present the models employed in our study: XGBoost, RandomForest and LightGBM. Subsequently, we analyze the causal inference methods and probability distribution estimations techniques, followed by the data augmentation approach. The chapter continues with a presentation of model calibration strategies and evaluation metrics and concludes with the results analysis.

6.1 XGBoost

eXtreme Gradient Boosting (XG-BOOST) was introduced by Chen and Guestrin in 2016 [16], emerging as a key innovation in the machine learning landscape. The algorithm was initially developed at the University of Washington, deriving from research focused on finding a more efficient and scalable implementation of gradient-boosting machines. Its development was motivated by the need for a system that could handle large-scale data while maintaining computational efficiency and prediction accuracy. After a machine learning competition hosted by Kaggle in 2015 [41], the XGBoost algorithm gained significant recognition, being used in 17 out of 29 winning solutions.
XGBoost brought notable advancements over traditional gradient-boosting implementations, introducing novel techniques for handling sparse data, tree pruning, and parallel computing. The system was designed to be highly portable and adaptable, also being able to run in different distributed environments while preserving performance [16].

6.1.1 Applications in Time Series Forecasting

The unique challenges presented by time series forecasting can be addressed by XGBoost which is well-equipped for the task. The ability of the algorithm to capture non-linear relationships and handle complex interactions between features makes it suited for temporal data analysis [68]. In the context of time series forecasting, XGBoost has demonstrated successful achievements across various domains, from financial market prediction [55] to climate modeling [27].
Recent applications have shown the effectiveness of XGBoost in conflict prediction and

in the social science domains. In particular, [13] employed XGBoost for conflict prediction in Africa, achieving good performance. The study showed that the algorithm is capable of processing multiple temporal features while accounting for geographical and socio-economical factors.
In the context of time series forecasting, several studies have demonstrated the effectiveness of the XGBoost algorithm:

- [54] showed the ability of XGBoost to predict stock market trends, well handling high-frequency trading data.
- In climate science, Kumar et al. [31] employed XGBoost for precipitation forecasting, improving the resulting accuracy over traditional time series models.
- [33] applied XGBoost to power load forecasting, achieving state-of-the-art results and high computational efficiency.

6.1.2 Technical Framework and Mathematical Foundations

The architecture behind XGBoost is built upon the principles of gradient boosting, with some key innovations that improve its performance and efficiency. The algorithm consists of constructing an ensemble of weak learners, typically decision trees, following an iterative process that minimizes a regularized objective function.

Mathematical Formulation

The core of the prediction model can be expressed through the following equation:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{6.1}$$

where $\mathcal{F}$ represents the space of regression trees, f_k corresponds to an independent tree structure, and K is the total number of trees.

The objective function being optimized is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6.2}$$

where l represents the training loss function and Ω is the regularization term.

The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{6.3}$$

where T represents the number of leaves in the tree, w represents the leaf weights, and γ and λ are regularization parameters.

Key Algorithmic Innovations

Part of the performance achieved by XGBoost is coming from different technical innovations incorporated into it:

- **Sparsity-Aware Split Finding**: the algorithm implements an optimized approach for handling missing values and sparse data through:

$$\text{gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{6.4}$$

- **Weighted Quantile Sketch**: XGBoost employs a weighted quantile sketch algorithm to efficiently handle continuous features, maintaining a balance between accuracy and computational efficiency.

- **Block Structure for Parallel Learning**: the algorithm utilizes a block structure to enable parallel computation:
  - feature values are first sorted and then stored in block structures
  - blocks can be distributed across machines for parallel processing
  - memory usage is optimized through the implementation of cache-aware access patterns

Advantages in Time Series Context

The architecture of XGBoost provides several tools that can be relevant for achieving good performance in the context of time series forecasting:

- **Feature Importance Quantification**: the model provides built-in mechanisms for evaluating feature importance, which is critical for understanding temporal dependencies:

$$\text{Importance}(f_j) = \sum_{k=1}^{K} \sum_{t \in T_k} \text{gain}(s_t) \cdot 1(v(s_t) = j) \tag{6.5}$$

- **Regularization Mechanisms**: the combination of L1 and L2 regularization helps prevent overfitting, which is particularly important when dealing with seasonal patterns and temporal dependencies.

- **Computational Efficiency**: XGBoost efficiently handles large datasets, making it particularly suitable for extended time series with multiple features thanks to the approximate tree learning algorithm, the cache-aware prefetching, and the out-of-core computation.

6.1.3 Implementation Considerations

In order to achieve optimal performance in time series forecasting, four key hyperparameters require particular care in the tuning phase:

- `max_depth`: controls the maximum depth of trees (range: 3-5), which is significant for capturing temporal dependencies
- `learning_rate`: Affects the step size in gradient descent (range: 0.01-0.1)
- `n_estimators`: determines the number of boosting rounds (range: 100-300). If set too low, the exposure to underperformance is inevitable, while if it is too high, there is a high risk of overfitting
- `subsample`: controls the fraction of samples used for tree building (range: 0.8-0.9)
- `colsample_bytree`: controls the fraction of features used when constructing each tree (range: 0.8-0.9), helping prevent overfitting
- `min_child_weight`: minimum sum of instance weight needed in a child node (range: 1-3), enhancing model robustness

6.2 LightGBM

Microsoft Research developed the Light Gradient Boosting Machine (LightGBM) in 2017 [29], which led to a significant advancement in gradient boosting frameworks. The algorithm emerged as a response to the computational challenges faced by existing gradient boosting models when addressing large-scale datasets. Its development was driven by the need for a more efficient approach to managing machine learning problems on an industrial scale while maintaining high accuracy. The framework gained rapid adoption in the data science community due to its speed and resource efficiency, particularly when implemented for tasks operating with large-scale temporal datasets [52].
The "Light" term, in the name of the algorithm, refers to its efficiency in handling training samples and feature behaviors, making it suited for large-scale applications [29]. Since its introduction, LightGBM has been widely employed in many industrial applications involving time series forecasting and real-time prediction systems.

6.2.1 Applications in Time Series Forecasting

The architecture of LightGBM offers some unique advantages for time series forecasting applications. The ability to handle large-scale temporal data while maintaining computational efficiency has made it extremely valuable in domains requiring real-time or near-real-time predictions [28].
In the context of time-dependent predictions, several studies have demonstrated the effectiveness of implementing the LightGBM model:

- In the context of link prediction, [6] employed LightGBM, achieving good performance even when handling temporal dependencies and categorical features.

- [17] utilized LightGBM for haze risk assessment, showing improved computational efficiency while maintaining prediction accuracy.
- In financial forecasting, [34] demonstrated the high capability of LightGBM in processing high-frequency trading data with high speed and accuracy.

The success of the algorithm in time series applications can be attributed to two main characteristics:

1. **Efficient Feature Handling**: the automatic handling of categorical features through optimal binning helps the algorithm to efficiently prepare them for the following steps. In the context of time series data, LightGBM efficiently processes the temporal lag features and can automatically manage missing values in time series data.
2. **Scalability for Large Temporal Datasets**: the algorithm is highly scalable thanks to the capability of processing multiple time series in parallel, the efficient memory utilization, especially for long sequence data, and the fast updating of the model.

6.2.2 Technical Framework and Mathematical Foundations

LightGBM introduces two significant innovations, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), that distinguish it from other gradient-boosting algorithms.

Mathematical Formulation

The core of the prediction model follows the gradient boosting framework:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad f_t \in \mathcal{F} \tag{6.6}$$

where $\mathcal{F}$ represents the space of regression trees, and T is the number of iterations.

The objective function is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{6.7}$$

Gradient-based One-Side Sampling (GOSS)

GOSS represents an approach for the selection of the samples for gradient boosting. The algorithm retains all instances with large gradients and randomly samples instances with small gradients:

$$\tilde{g}_i = \begin{cases} g_i & \text{if } |g_i| \geq \text{top}_a \\ \frac{g_i}{b} & \text{if } |g_i| < \text{top}_a \text{ and selected} \end{cases} \tag{6.8}$$

where top_a represents the threshold for large gradients, and b is the sampling ratio for small gradients.

Exclusive Feature Bundling (EFB)

EFB addresses the challenge of high-dimensional feature spaces by grouping mutually exclusive features. The bundling process can be formulated as a graph coloring problem:

$$\text{Bundle}(F_1, F_2) = \begin{cases} 1 & \text{if conflict}(F_1, F_2) \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{6.9}$$

where θ is the conflict threshold for feature bundling.

6.2.3 Key Innovations and Advantages

1. **Leaf-wise Growth Strategy**: differently from the traditional level-wise tree growth, LightGBM employs a leaf-wise approach shown in the following equation:

$$\text{Gain}_{\text{split}} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] \tag{6.10}$$

2. **Feature Discretization**: LightGBM implements histogram-based feature discretization which is responsible for reducing memory usage by converting continuous features into discrete bins, and accelerating training by computing gradients on bins rather than individual values. It also enables efficient parallel processing.

6.2.4 Implementation Considerations

For optimal performance in time series forecasting, the tuning of some parameters is fundamental and requires particular consideration. Two types of parameters can be distinguished: core parameters and time series specific parameters.

Core Parameters

- `num_leaves`: Maximum number of leaves in one tree (range: 31-50)
- `max_depth`: Maximum depth of the tree (range: 5-10)
- `learning_rate`: Step size of each iteration (range: 0.01-0.1)
- `n_estimators`: Number of boosting iterations (range: 100-200)
- `colsample_bytree`: Fraction of features to be used in each iteration (range: 0.8-0.9)

Time Series Specific Parameters

- `subsample`: Fraction of data to be used for each iteration (range: 0.8-0.9)
- `min_child_samples`: Minimum number of records required in a leaf node (range: 20-50)

Optimization for Temporal Data

When applying LightGBM to time series forecasting, special attention should be paid to the feature engineering process and to the optimization of memory usage. The first is important when creating appropriate lag features, handling seasonal components, or integrating external temporal indicators. The memory optimization process follows the usage of the histogram-based algorithm for continuous features, the Efficient Feature Handling, and the application of Exclusive Feature Bundling.

6.3 Random Forests

Random Forests, introduced by Leo Breiman in 2001 [10], are an ensemble learning method that has maintained its relevance and effectiveness even in modern machine learning applications. The algorithm followed Breiman's earlier work on bagging predictors [9], combining the concepts of bagging with random feature selection. This innovation addressed the limitations of single decision trees while maintaining their interpretability and robustness.
Random Forests marked a significant advancement in ensemble methods, introducing a technique that reduces variance without increasing bias, which is a common challenge in machine learning models. Since their introduction, RFs have become a standard basis in predictive modeling, particularly valued for the stability and resistance to overfitting [64].

6.3.1 Applications in Time Series Forecasting

Random Forests were initially designed for standard classification and regression tasks, however through the remarkable adaptability intrinsic to them, the employment in time series forecasting problems was inevitable. The ability to capture non-linear relationships and manage high-dimensional feature spaces makes RFs suitable for temporal predictions [26].

In the context of conflict prediction and time series applications, several notable studies have demonstrated RF's effectiveness. [19] studied the performance of RFs for conflict prediction achieving better accuracy levels with respect to other machine learning and deep learning models, demonstrating the ability to handle multiple tempo-spatial features. [37] utilized RFs for political violence prediction, showing robust performance even in the presence of imbalanced temporal data.

The key advantages of Random Forests, when employed in time series data, are the temporal feature handling and the intrinsic stability of the ensemble. The first advantage

comes from the natural incorporation of lag features and the robustness to missing values in time series, while the stability is achieved from the reduced variance through averaging multiple trees, which causes a certain resistance to overfitting and to outliers.

6.3.2 Technical Framework and Mathematical Foundations

Random Forest constructs an ensemble of decision trees through two randomization processes: bootstrap sampling of instances (bagging) and random selection of features at each split.

Mathematical Formulation

The core prediction model for regression can be expressed as:

$$\hat{f}_{\mathrm{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{6.11}$$

where B is the number of trees, and $T_b(x)$ represents the prediction of the b-th tree.

The variance of the forest estimate can be decomposed as:

$$\mathrm{Var}(\hat{f}_{\mathrm{RF}}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{6.12}$$

where ρ represents the correlation between trees, and σ^2 is the variance of individual trees.

Tree Construction Process

For each tree in the forest:

1. **Bootstrap Sampling**: draw a bootstrap sample of size n from the training data with probability:
$$P(X_i \text{ selected}) = 1 - (1 - \frac{1}{n})^n \approx 0.632 \tag{6.13}$$

2. **Random Feature Selection**: at each node, randomly select m features from the total p features. For the regression task $m \approx p/3$, while for classification $m \approx \sqrt{p}$.

3. **Node Splitting**: for regression trees, the split criterion is typically the reduction in variance given by:
$$\Delta I = \frac{1}{N_m}\sum_{i\in N_m}(y_i - \bar{y}_m)^2 - \sum_{j\in\{L,R\}}\frac{N_j}{N_m}\frac{1}{N_j}\sum_{i\in N_j}(y_i - \bar{y}_j)^2 \tag{6.14}$$
where N_m represents the number of instances at node m.

6.3.3 Key Innovations

A key innovation introduced by Random Forests is the **Out-of-Bag Error Estimate** which provides an unbiased estimate of the generalization error and is given by:

$$\text{OOB}_{\text{error}} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_{\text{OOB}}(x_i)) \tag{6.15}$$

Another main innovation in the algorithm is the **Variance Importance Measure** which comprehends two methods for assessing feature importance: *Mean Decrease in Impurity (MDI)* and *Mean Decrease in Accuracy (MDA)* given by the following equations.

$$\text{MDI}(X_j) = \sum_{t=1}^{B} \sum_{m \in T_b} p(m)\Delta i(s_m, j) \tag{6.16}$$

$$\text{MDA}(X_j) = \text{OOB}_{\text{error}}(\tilde{X}_j) - \text{OOB}_{\text{error}}(X_j) \tag{6.17}$$

6.3.4 Implementation Considerations for Time Series

For the implementation of RFs in time series prediction tasks, some important considerations must be considered.
Another key aspect to consider is the optimization of the hyperparameters, in particular:

- `n_estimators`: number of trees in the forest (range: 100-200)
- `max_depth`: maximum depth of each tree (range: 5-10)
- `min_samples_split`: minimum samples required to split a node (range: 2-5)
- `min_samples_leaf`: minimum samples required in a leaf node (range: 1-2)
- `max_features`: number of features to consider for best split (options: 'sqrt', 'log2')

Finally, although more implicit than the others, performance optimization is critical when considering large datasets. A parallel processing implementation can significantly speed up the predictions.

6.3.5 Advantages for Conflict Prediction

A key advantage of RFs is the **robustness** to noise, as the approach effectively handles measurement errors in conflict data, resists outliers in temporal patterns, and remains stable even with missing data. Additionally, they excel in the detection of feature interactions, capturing complex relationships between predictors, accommodating non-linear temporal dependencies, and efficiently processing both categorical and numerical features. Finally, the algorithm enhances **interpretability** by providing clear rankings of feature importance, facilitating the understanding of predictor effects through partial dependence plots, and enabling the assessment of variable interactions.

6.4 Causal Inference

In the context of conflict prediction and fatality forecasting, understanding the underlying causal relationships between socio-political factors is fundamental. Causal inference, as defined by Pearl [47], is a branch of machine learning that aims to answer causal queries from data with minimal experimentation and assumptions. This section explores how causal inference methods can enhance early warning systems for conflict prediction by enabling researchers to reason about questions such as "What would be the fatality levels if a country entered a state of conflict?"

6.4.1 Motivation and Background

Machine learning models heavily rely on large quantities of high-quality data to perform well. However, in real-world scenarios, particularly in conflict prediction, data often lacks in both quantity and quality. Many datasets fail to cover the entire distribution over which the model operates, leading to poor generalization. While foundational models like SAM [30], GPT [11], and Llama [58] have partially addressed this challenge in computer vision and natural language processing, the time-series domain presents unique challenges.

In time series forecasting for conflict prediction, deep learning methods, including transformers, have shown limited effectiveness [23]. Traditional time series models like XGBoost, while efficient to train, still face generalization challenges when historical data fails to capture all possible conflict scenarios or domain shifts. This necessitates an alternative approach to the generalization problem, focusing on data generation rather than model complexity.

6.4.2 Foundations of Causal Modeling

Graphical models serve as the primary tool for executing causal queries, as they can compactly capture relationships between features and potentially explain them through mathematical functions [32]. In our context, these models help us understand how different factors like political stability, demonstrations, and strategic developments causally influence fatality levels.

6.4.3 Domain Adaptation through Causal Intervention

A fundamental challenge in conflict prediction is the handling of domain shifts, particularly when a region transitions from relative peace to conflict. Traditional machine learning approaches struggle with these shifts because they violate the common assumption that training and test data come from the same distribution. Our approach addresses this by dividing the records into two domains based on fatality levels, after experimenting with alternative splitting criteria such as political violence and quiet period indicators, which proved less effective.

Consider two domains representing different conflict states:

$$X_{c,d_1} \rightarrow X_{c,d_2} \tag{6.18}$$

where c represents a country and d_1, d_2 represent different domains characterized by low and high fatality levels respectively. The split between domains is determined using a dynamic threshold based on the 75th percentile of fatalities for each country, with additional balancing mechanisms to ensure sufficient samples in both domains.

Our solution leverages the data of each country, divided into these two domains, to generate synthetic samples that maintain the country-specific characteristics while exploring different conflict intensities. This approach allows the model to learn from both peaceful and tumultuous periods while preserving the unique patterns of each region.

6.5 Probability Distribution Estimation

To perform effective data augmentation in our two-domain framework, we employ two complementary approaches: Kernel Density Estimation (KDE) and Natural Neighbor Interpolation (NNI). Each method has the advantage of preserving the statistical properties of the original data while generating meaningful synthetic samples.

6.5.1 Kernel Density Estimation

In our implementation, KDE employs a Gaussian kernel with RBF length scale of 1.0. The process begins with standard scaling of the data to normalize feature distributions. Then sample weights are computed using the RBF kernel, followed by the generation of new samples with controlled Gaussian noise at scale 0.1 to increase variety while maintaining data structure. With this approach we make sure that the generated samples maintain the characteristics of the original data distribution and the temporal dependencies, which are crucial for conflict prediction.

6.5.2 Natural Neighbor Interpolation

Our NNI implementation allows to diversify sample generation through multiple complementary strategies. The process combines direct interpolation between randomly selected point pairs, base point modification with scaled Gaussian noise, and feature-wise mixing with random interpolation weights. This approach employing multiple strategies helps to maintain local data structures while, at the same, introducing enough variation in the generated samples.

6.5.3 Implementation Considerations

The implementation processes each country independently to maintain its unique characteristics, employing forward and backward filling for missing values. The temporal structure is preserved through weekly frequency spacing, with a fixed augmentation ratio of 2 that effectively doubles the available samples for each domain. With this approach we make sure that both the country-specific patterns and consistent temporal ordering in the generated features are preserved.

6.6 Causal Data Augmentation

Building upon the probability distribution estimation methods described above, our data augmentation process follows three main steps:

1. **Domain Splitting**: For each country, we split the data into low and high fatality domains using the 75th percentile threshold, with dynamic adjustment to ensure sufficient samples in each domain;

2. **Parallel Generation**: We generate new samples using both KDE and NNI methods independently, preserving the temporal structure and country-specific patterns;

3. **Data Integration**: The generated samples are combined with the original data, maintaining temporal ordering and ensuring consistent weekly spacing.

The process is applied separately for each country in the dataset, resulting in two augmented datasets (KDE-based and NNI-based) which doubles the size of the original data while maintaining its essential characteristics and temporal structure.

6.7 Model Calibration

Model calibration is crucial in conflict prediction systems where decision-making relies heavily on the confidence levels associated with predictions. A well-calibrated model provides reliable probability estimates, essential for stakeholders to make informed decisions about preventive actions.

6.7.1 Conformal Prediction Framework

Conformal prediction [53] represents the state of the art in predictive model calibration. The method relies on the exchangeability assumption, which posits that given a set of N outcomes, any ordering of these outcomes is equally likely. Under this assumption, conformal prediction uses a calibration set to generate prediction intervals with theoretical guarantees.

In our implementation, we employ Inductive Conformal Prediction (ICP) through the IcpRegressor framework, using absolute error as the non-conformity measure. This choice allows us to handle regression problems while maintaining computational efficiency.

However, two significant challenges arise in our context:

1. The exchangeability assumption often fails in time series data, where the order of observations inherently matters

2. Traditional conformal prediction does not account for covariate shift, a common occurrence in conflict prediction where data distributions can change significantly

6.7.2 Adaptation for Distribution Shifts

To address these challenges, we follow an approach inspired by Tibshirani et al. [57], who propose using likelihood ratios to adapt conformal prediction under covariate shift. The method works by weighing samples to match the calibration set distribution to the target distribution:

$$w(x) = \frac{p_{\text{target}}(x)}{p_{\text{source}}(x)} \tag{6.19}$$

While this approach theoretically addresses distribution shift, it requires knowledge of the likelihood ratio, which is often difficult to estimate in practice. For this reason, we do not explicitly compute these likelihood ratios, however, our approach achieves similar objectives through data augmentation. By generating synthetic calibration data that spans both peaceful and tumultuous periods, we effectively re-weight the calibration set to better match target distributions we might encounter in real data.

6.7.3 Causal Data Augmentation for Calibration

Our contribution applies data augmentation to the calibration process. We:

1. Generate synthetic calibration data using our CDA framework

2. Apply conformal prediction on the augmented calibration set

3. Evaluate the results in different scenarios

Its practical implementation relies on a careful segmentation of our dataset. We partition the data temporally into three sets: a training set for model fitting, a validation set that serves as our calibration set, and a test set for final evaluation. This temporal splitting is essential for respecting the time-series nature of conflict data and allows us to evaluate the model performance under realistic conditions where we must predict future events using only past information.

More specifically, our implementation uses the validation set as the calibration set, following inductive conformal prediction principles. For each country, we apply separately KDE and NNI based augmentation to enrich the calibration data . This approach helps ensure robust calibration across different conflict intensities not present in the original data.

6.8 Evaluation Metrics

The evaluation of conflict prediction models requires metrics that can capture general predictive accuracy and specific causes to detect critical events.

6.8.1 Base Performance Metrics

The first metric we employ for the model training in the Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6.20}$$

Where y_i is the real number of fatalities and $\hat{y}_i$ is the predicted value.

6.8.2 Spike Detection Metrics

Given the critical nature of conflict escalation, we introduce two specialized metrics: Spike Precision (SP) and Spike Recall (SR). Initially, we experimented with an adaptive threshold approach that categorized countries based on their historical 90th percentile of monthly fatalities, with different parameters basef on conflict intensity: high-intensity (>500 fatalities), medium-intensity (150-500), and low-intensity (<150). However, this approach proved to be problematic as it introduced excessive complexity and made cross-country comparisons difficult, potentially biasing the model's behavior. Furthermore, the varying thresholds made it challenging to establish a general guideline.

For this reason, we decided to opt for a simpler, more generalizable approach with two parameters that vary based on the prediction horizon rather than country-specific characteristics (for which we already account through the 95th percentile fatality truncation as discussed):

- **Spike Threshold (α)**: Defines the minimum number of fatalities to classify an event as a spike
- **Spike Tolerance (β)**: Defines the acceptable prediction error tolerance for spike detection

Formally, for a prediction $\hat{y}$ and ground truth y, we classify a correct spike detection when:

$$|y - \hat{y}| < \beta \text{ and } y > \alpha \tag{6.21}$$

We then define:

$$\text{SP} = \frac{\text{Correctly Predicted Spikes}}{\text{Total Predicted Spikes}} \tag{6.22}$$

$$\text{SR} = \frac{\text{Correctly Predicted Spikes}}{\text{Total Actual Spikes}} \tag{6.23}$$

For instance, in our experiments with a 4-week prediction horizon we set $\alpha = 50$ fatalities and $\beta = 20$ fatalities, based on domain expertise and operational requirements. These parameters are adjusted proportionally for different prediction horizons, ensuring consistent spike detection capabilities across different temporal scales while maintaining a uniform approach across countries.

6.8.3 Evaluation on Data Augmentation Methods

In order to measure the quality of the augmentation techniques, we employed two distribution similarity metrics: the Wasserstein distance and the Jensen-Shannon divergence. The first measures the minimum "cost" to transform one distribution into another, while the other measures the similarity between probability distributions.

Figure 6.1 shows the comparison between KDE and NNI methods across all features. The points above the diagonal line indicate the features where NNI outperforms KDE, while the points below suggest better performances of KDE. The analysis reveals that NNI consistently obtains lower divergence scores, which is particularly evident in the Wasserstein distance plot where most points lie below the diagonal. This suggests that NNI preserves more faithfully the original data distributions compared to KDE.

The distribution preservation is particularly evident in key features such as `goldstein_weighted` and `media_pressure`, as shown in Figure 6.2. For the `goldstein_weighted` feature, NNI demonstrates high capability in maintaining the shape of the distribution, with a peak around -5 and a rapid decline on both sides. The KDE method, while still capturing the general shape, shows a more smoothed distribution with a lower peak and wider spread. The `media_pressure` distributions reveal similar patterns, with NNI closely following the original distribution's shape and peak around value 5, while KDE exhibits a more spread-out distribution with a lower peak intensity. This behavior is consistent with the theoretical properties of both methods: NNI tends to preserve local structures and sharp features in the data, while KDE naturally introduces some smoothing due to its kernel-based approach.

The analysis reveals that while both methods maintain the overall structure of the data, NNI shows particular strength in preserving the detailed characteristics of the distributions.

This result ca be explained from NNI's ability to maintain local density patterns without introducing excessive smoothing. KDE, while still effective, in comparison tends to produce slightly more diffuse distributions, which could be less suitable for features where sharp transitions are present.

6.9 Results Analysis

In this section, we present a comprehensive analysis of the improvements achieved through data augmentation techniques (KDE and NNI) compared to the original dataset. The obtained results are analyzed from multiple perspectives: general improvements, consistency of improvements, model-specific analysis, and country-level analysis.

6.9.1 General Improvements

Table 6.1 shows the overall improvements achieved by each method across different metrics. NNI reached the best performance, specifically regarding spike detection capabilities, with a remarkable **13.1%** improvement in Spike_F1 score. In particular, both methods

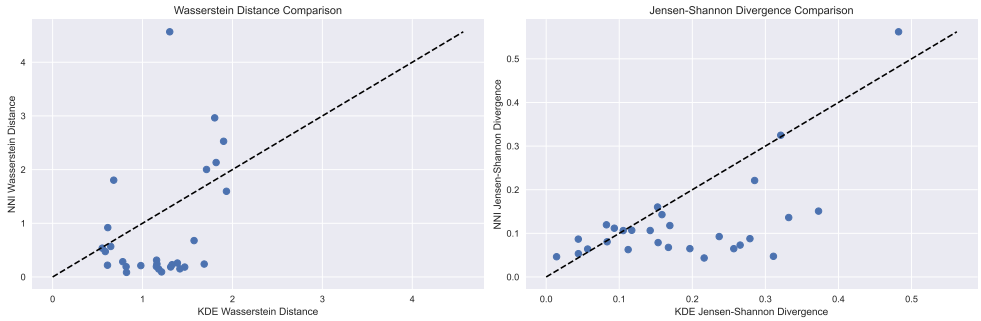

Figure 6.1. Comparison of augmentation methods using Wasserstein distance (left) and Jensen-Shannon divergence (right). Points below the diagonal line indicate features where NNI achieves better performance than KDE.

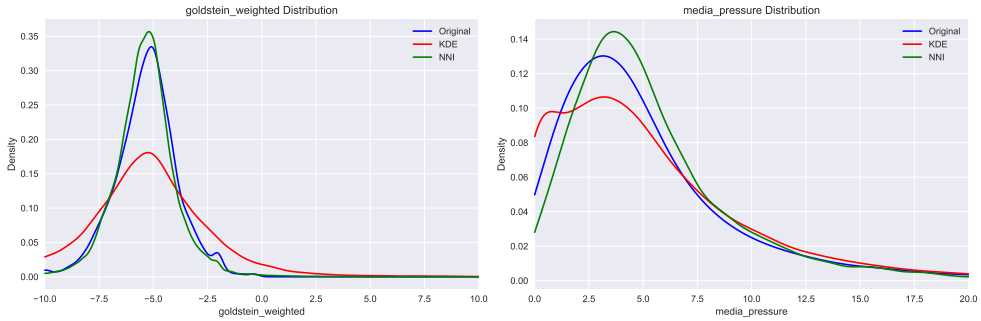

Figure 6.2. Distribution comparison of key features across original and augmented datasets. The plots show how both KDE and NNI preserve the original distributions, with NNI (green) typically achieving closer alignment to the original data (blue) than KDE (red).

show positive improvements in error metrics (MAE and RMSE), with NNI achieving better results (**6.6%** and **5.9%** respectively).

6.9.2 Improvement Consistency

Table 6.2 presents the consistency of improvements across different metrics. NNI shows remarkable consistency in error metrics, with **80%** of cases showing improvement in both MAE and RMSE. Moreover, it also achieves the highest consistency in Spike_Precision at **83.3%**.

6.9.3 Model-Specific Analysis

Table 6.3 highlights the presence of patterns in model-specific improvements. LightGBM shows the most substantial improvements with both methods, achieving a remarkable

Table 6.1. General performance improvements, ordered by improvement percentage, compared to the original dataset

Metric	Method	Absolute Improvement	Improvement %
Spike_F1	NNI	0.035	**13.100**
MAE	NNI	**5.827**	**6.677**
RMSE	NNI	**6.818**	**5.967**
Spike_F1	KDE	0.015	5.615
Spike_Precision	NNI	**0.033**	4.795
RMSE	KDE	3.990	3.492
MAE	KDE	2.603	2.982
Spike_Precision	KDE	-0.005	-0.727
Spike_Recall	NNI	-0.005	-2.049
Spike_Recall	KDE	-0.009	-3.619

Table 6.2. Percentage of cases showing improvement by metric and method

Method	Metric	% Positive Improvements
KDE	MAE	60.00
	RMSE	60.00
	Spike_F1	58.33
	Spike_Precision	50.00
	Spike_Recall	50.00
NNI	MAE	**80.00**
	RMSE	**80.00**
	Spike_F1	58.33
	Spike_Precision	**83.33**
	Spike_Recall	41.67

15.3% average improvement with KDE and **14.5%** with NNI. XGBoost shows contrasting results between methods: although NNI achieves positive improvements, KDE shows consistent degradation.

6.10 Calibration Results and Analysis

The conformal prediction calibration was performed on multiple countries in the Sahel region, using both augmented and original datasets with a 95% confidence level. The results demonstrate varying degrees of effectiveness across different countries, with an overall average coverage of 59.7% and an efficiency of 24.6%. To illustrate the model's performance across different conflict scenarios, we present three representative cases: Central African Republic (CAR), Niger, and Mali (Figures 6.3, 6.4, and 6.5).

The analysis reveals changing model performance across different conflict dynamics. The Central African Republic presents a case of decreasing conflict intensity, where the

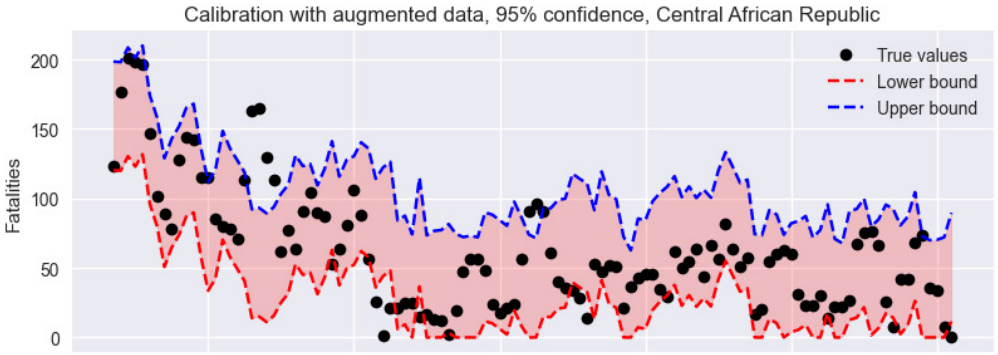

Figure 6.3. Calibration results for Central African Republic showing a decreasing trend in conflict intensity. The prediction intervals (average width 102.41) and high coverage (94%) demonstrate the model's ability to adapt to declining conflict patterns. Note how the bounds effectively capture the gradual reduction in fatalities over time.

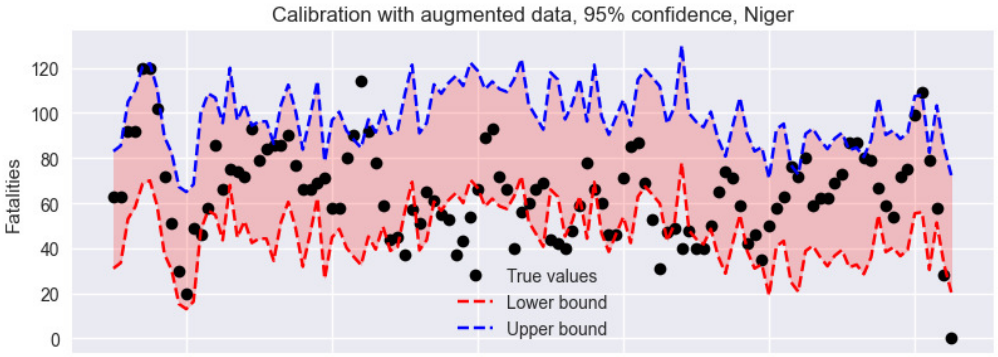

Figure 6.4. Calibration results for Niger illustrating medium-intensity conflict prediction. The moderate prediction intervals (average width 21.66) and balanced coverage between peace (20%) and conflict (57%) periods show how the model handles intermediate conflict scenarios. True values (black dots) demonstrate significant variability within the prediction bounds.

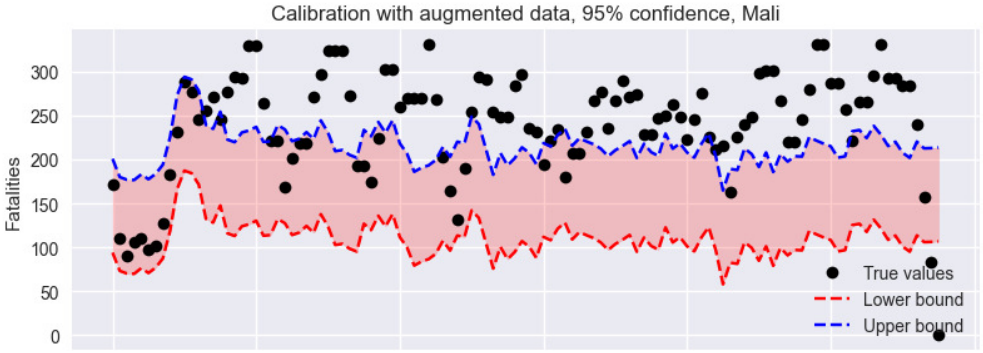

Figure 6.5. Calibration results for Mali demonstrating sustained high-intensity conflict prediction. The prediction intervals (average width 102.87) capture a consistent level of elevated fatalities (200-300 range), showing how the model adapts to persistent conflict scenarios with relatively stable bounds despite high fatality numbers.

Table 6.3. Average percentage improvements by model and augmentation method

Model	Method	Mean	Std	Min	Max
LightGBM	KDE	**15.320**	15.572	5.248	**41.843**
	NNI	**14.494**	13.554	5.926	38.178
RandomForest	KDE	-2.848	10.836	-19.445	5.985
	NNI	-1.042	11.337	-18.475	7.433
XGBoost	KDE	-5.388	4.431	-11.191	-0.677
	NNI	5.353	5.360	0.794	14.185

model achieves excellent coverage (94%) while adapting its prediction intervals to a declining trend in fatalities. The high coverage during both peace and conflict periods (98% and 90.8% respectively) demonstrates the model's ability to track gradual transitions in conflict intensity.

Niger represents an intermediate scenario with moderate but variable conflict levels. The model maintains reasonable coverage (47.4%) with relatively narrow prediction intervals (width: 21.66), showing interesting differences between peace and conflict periods (20% and 57% coverage respectively). This case highlights the challenges in balancing prediction accuracy across varying intensity levels within the same region.

Mali presents a distinct pattern of sustained high-intensity conflict, where the model maintains consistent prediction intervals (width: 102.87) for persistently elevated fatality numbers. This case is particularly noticeable as it demonstrates the model's behavior in scenarios of sustained conflict, rather than sporadic outbreaks or peaceful periods. It achieves 36.2% coverage despite the challenging nature of persistent high-intensity predictions.

Several other countries exhibited notable performance characteristics, such as Chad (91.4% coverage), South Sudan (90.5%), and Burkina Faso (20.7%), with different interval widths. The analysis revealed interesting patterns in peace-conflict dynamics, with some countries showing marked differences in prediction accuracy between peaceful and conflict periods. For instance, Chad achieved 100% coverage during peace periods but only 33.3% during conflicts.

When attempting to reduce interval widths by adjusting the significance level to 90%, we observed a trade-off between coverage and efficiency: while the coverage decreased from 59.7% to 49.6%, the efficiency improved from 24.6% to 27%. However, given the critical nature of conflict prediction, maintaining higher coverage with wider prediction intervals was considered preferable despite the slightly lower efficiency. For instance, this decision can be justified by the model's performance in cases like Mali, where consistent coverage of sustained high-intensity conflicts is crucial for early warning systems.

Chapter 7

Conclusions and Future Work

This thesis has explored a new approach for conflict prediction. The results demonstrate that our model achieves good improvements in both general prediction accuracy and spike detection capability. The best finding is the effectiveness of NNI for data augmentation, which showed a 13.1% improvement in F1 score and approximately 6.7% improvement in Mean Absolute Error compared to baseline models. Among the machine learning models tested, LightGBM demonstrated the most substantial improvements when combined with our augmentation techniques, achieving up to 15.3% better performance across various metrics. The calibration framework revealed a good model ability to adapt to different conflict scenarios, from declining intensity to sustained high-level conflicts, while maintaining reliable but probably wide prediction intervals.

Despite the interesting findings, it can be improved under different aspects. The model's performance showed some variability across different regions, particularly in cases of rapid conflict escalation, which is one of the main goals of a conflict early warning system. Additionally, while our data augmentation techniques generally improved model performance, their effectiveness varied across different countries and conflict intensities, suggesting that it can be further optimized.

7.1 Future Work

There can be several directions for future research from this work:

- **Real-time News Data Approach**: Developing more sophisticated techniques for processing a huge amount of online news articles and social media data, since they are real-time data sources and can capture emerging informations immediately.
- **Hybrid Augmentation Approaches**: Develop a dynamic framework that combines KDE, NNI or other augmentation methods.
- **Improved regional Customization**: Implementing region-specific augmentation parameters and adaptive rates based on local characteristics could be interesting in order to enhance local model performance, which is one of the main issues.

- **Feature Engineering**: Investigating additional feature engineering approaches, specifically tailored for conflict prediction.

Bibliography

[1] *Perilous Desert: Insecurity in the Sahara*. Carnegie Endowment for International Peace, 2013.

[2] Use of early warning systems in western africa to combat terrorism. 2019.

[3] Global terrorism index 2024. 2024. Accessed 2024.

[4] ACLED. Acled data codebook. 2023.

[5] Mohammad Al-Saidi, S.A. Saad, and Nadir Elagib. From scenario to mounting risks: Covid-19's perils for development and supply security in the sahel. *Environment, Development and Sustainability*, 25, 04 2022.

[6] Asia Mahdi Naser Alzubaidi. Lightgbm for link prediction based on graph structure attributes. *Education*, 2020, 2016.

[7] Marco Barreno, Blaine Nelson, Anthony Joseph, and J. Tygar. The security of machine learning. *Machine Learning*, 81:121–148, 11 2010.

[8] Patrick Brandt and Todd Sandier. What do transnational terrorists target? has it changed? are we safer? *Journal of Conflict Resolution*, 54:214–236, 01 2010.

[9] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[10] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[13] Tim Busker, Bart van den Hurk, Hans de Moel, Marc van den Homberg, Chiem van Straaten, Rhoda A Odongo, and Jeroen CJH Aerts. Predicting food-security crises in the horn of africa using machine learning. *Earth's Future*, 12(8):e2023EF004211, 2024.

[14] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24, 04 2018.

[15] Emily Chen and Emilio Ferrara. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between ukraine and russia. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1006–1013, Jun. 2023.

[16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[17] Hongbin Dai, Guangqiu Huang, Huibin Zeng, and Rongchuan Yu. Haze risk assessment based on improved pca-mee and ispo-lightgbm model. *Systems*, 10(6):263, 2022.

[18] Daniel Eizenga. Long term trends across security and development in the sahel. (25), 2019.

[19] Felix Ettensperger. Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. *Quality & Quantity*, 54(2):567–601, 2020.

[20] Francis N. Okpaleke Ezenwa E. Olumba, Bernard U. Nwosu and Rowland Chukwuma Okoli. Conceptualising eco-violence: moving beyond the multiple labelling of water and agricultural resource conflicts in the sahel. *Third World Quarterly*, 43(9):2075–2090, 2022.

[21] Institute for Economics & Peace. Ecological threat register 2020: Understanding ecological threats, resilience and peace. September 2020. Accessed: Date Month Year.

[22] Sabrina Gabel, Lilian Reichert, and Christian Reuter. Discussing conflict in social media: The use of twitter in the jammu and kashmir conflict. *Media, War & Conflict*, 15(4):504–529, 2022.

[23] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1, 2024.

[24] Alessandra Giannini, Ramalingam Saravanan, and P Chang. Oceanic forcing of sahel rainfall on interannual to interdecadal time scales. *Science (New York, N.Y.)*, 302:1027–30, 11 2003.

[25] Håvard Hegre, Angelica Lindqvist-McGowan, Paola Vesco, Remco Jansen, and Malika Rakhmankulova. Forecasting armed conflict in the sahel: Forecasts for november 2021–october 2024. 2022.

[26] Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

[27] Ziyu Jia, Zhenhao Zhang, Yunxiang Cheng, Shinchilelt Borjigin, Zhijia Quan, et al. Grassland biomass spatiotemporal patterns and response to climate change in eastern inner mongolia based on xgboost model estimates. *Ecological Indicators*, 158:111554, 2024.

[28] Dongzi Jin, Yiqin Lu, Jiancheng Qin, Zhe Cheng, and Zhongshu Mao. Swiftids: Real-time intrusion detection system based on lightgbm and parallel intrusion detection mechanism. *Computers & Security*, 97:101984, 2020.

[29] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[31] Vijendra Kumar, Naresh Kedam, Ozgur Kisi, Saleh Alsulamy, Khaled Mohamed Khedher, and Mohamed Abdelaziz Salem. A comparative study of machine learning models for daily and weekly rainfall forecasting. *Water Resources Management*, pages 1–20, 2024.

[32] Steffen L Lauritzen. Causal inference from graphical models. *Monographs on Statistics and Applied Probability*, 87:63–108, 2001.

[33] Yahui Liu, Huan Luo, Bing Zhao, Xiaoyong Zhao, and Zongda Han. Short-term power load forecasting based on clustering and xgboost method. In *2018 IEEE 9th international conference on software engineering and service science (ICSESS)*, pages 536–539. IEEE, 2018.

[34] Jiehua Lv, Chao Wang, Wei Gao, and Qiumin Zhao. [retracted] an economic forecasting method based on the lightgbm-optimized lstm and time-series model. *Computational Intelligence and Neuroscience*, 2021(1):8128879, 2021.

[35] Karen Meijer, Rolien Sasse, and Judith Blaauw. The WPS approach: A design for an integrated, inclusive and informed approach to address water-related security risks. Working paper, Water, Peace and Security Partnership, 2023. Contributors: Jessica Hartog, Camille Marquette, Susanne Schmeier.

[36] Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data Society*, In press, 10 2016.

[37] David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103, 2016.

[38] Uzma Naz and Muhammad Subhan Saleem. Climate-induced vulnerabilities: Conflict and migration patterns in the sahel region of africa. *Pakistan Languages and Humanities Review*, 8(2):295–311, Jun. 2024.

[39] Paul Nemitz. Constitutional democracy and technology in the age of artificial intelligence. *SSRN Electronic Journal*, 01 2018.

[40] Jerome Nenger. The impact of climate change on forced migration in the sahel: Human rights perspective (nigeria as a case study). 20:31, 02 2024.

[41] Didrik Nielsen. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU, 2016.

[42] Elias Nkiaka, Robert Bryant, and Zongho Kom. Understanding links between water scarcity and violent conflicts in the sahel and lake chad basin using the water footprint concept. *Earth's Future*, 12, 02 2024.

[43] OECD, Sahel, and West Africa Club. *An Atlas of the Sahara-Sahel*. 2014.

[44] Office of the United Nations High Commissioner for Human Rights. The human rights impacts of climate change on migration in the sahel. 2021.

[45] Intergovernmental Authority on Development (IGAD). Conflict early warning and response mechanism (cewarn). https://cewarn.org/.

[46] European Parliament and Council of the European Union. Regulation (eu) 2016/679

of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, L 119:1–88, 2016.

[47] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

[48] Jonathon Penney. Internet surveillance, regulation, and chilling effects online: a comparative case study. *Internet Policy Review*, 6, 05 2017.

[49] Akah Pius Odey and Osmond Otora. An assessment of ecowas early warning and crises respond network(ecowarn) and transborder criminality in west africa, 1999-2021. 15:2022, 08 2023.

[50] UNDP (United Nations Development Programme). Human development report 1990. *UNDP (United Nations Development Programme)*, 1990.

[51] Neil Richards and Woodrow Hartzog. Taking trust seriously in privacy law. *SSRN Electronic Journal*, 01 2015.

[52] Eray Sevgen and Saygin Abdikan. Classification of large-scale mobile laser scanning data in urban area with lightgbm. *Remote Sensing*, 15(15):3787, 2023.

[53] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[54] Priyanka Sharma and Mayank Kumar Jain. Stock market trends analysis using extreme gradient boosting (xgboost). In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 317–322. IEEE, 2023.

[55] LEI Shimin, XU Ke, Yizhe Huang, and SHA Xinye. An xgboost based system for financial fraud detection. In *E3S Web of Conferences*, volume 214, page 02042. EDP Sciences, 2020.

[56] Niklas Stoehr, Lucas Hennigen, Josef Valvoda, Robert West, Ryan Cotterell, and Aaron Schein. An ordinal latent variable model of conflict intensity. 10 2022.

[57] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[59] UNDP. Human development index (hdi).

[60] United Nations Development Programme. United nations development programme. 2023.

[61] United Nations Development Programme. Human Development Report 2022-23: Breaking the Gridlock: Reimagining Cooperation in a Polarized World. Accessed via Wikipedia, March 2024. Human Development Reports 288–292.

[62] United Nations Economic Commission for Africa. *Economic Report on Africa 2017: Urbanization and Industrialization for Africa's Transformation*. 2017. Accessed: 15 October 2024.

[63] ViEWS Project. Views - early warning system model documentation: fatalities002,

2024.

[64] Huazhen Wang, Fan Yang, and Zhiyuan Luo. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC bioinformatics*, 17:1–18, 2016.

[65] Claire Wardle and Hossein Derakhshan. *INFORMATION DISORDER : Toward an interdisciplinary framework for research and policy making Information Disorder Toward an interdisciplinary framework for research and policymaking*. 09 2017.

[66] World Bank. Country climate and development report, 2022.

[67] World Bank. World bank databank indexes, 2023.

[68] Xinmeng Zhang, Chao Yan, Cheng Gao, Bradley Malin, and You Chen. Xgboost imputation for time series data. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–3. IEEE, 2019.